MACHINE LEARNING AND OPTIMIZATION FOR HEALTHCARE AND ENERGY SYSTEMS

by

Wei Jiang

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland

May 2018

Abstract

Healthcare and energy systems provide critical service to our society. Recent advancement in information technology has enabled these systems to keep retrieving and storing data. In this dissertation, we used machine learning, optimization techniques, and data from healthcare and energy systems to build predictive models and discover new knowledge to guide decision-making and improve the efficiency and sustainability of these systems. We also used optimization techniques to improve the efficiency of hyperparameter tuning for machine learning algorithms. Specifically, we built a dynamic daily prediction model for predicting heart failure patients' 30-day readmission risk. We built a prediction model to predict xerostomia (dry mouth) for head and neck cancer patients treated with radiotherapy and identified the influence pattern of radiation dose across head and neck on xerostomia. Using an economic equilibrium model combined with optimization techniques for calibration, we built the first global trade model for wood chip and analyzed how local renewable energy policy in the United States could affect the global wood chip trade and lead to deforestation in other world regions. Finally, we created a new method for tuning the hyper-parameter for support vector machines by solving the problem as a bilevel optimization

problem using stochastic gradient descent combined with dual coordinate descent method. We showed that the new method is more efficient than ad hoc empirical approaches. In summary, we demonstrated how machine learning and optimization techniques can improve the efficiency of healthcare and energy systems, and how optimization techniques can advance machine learning algorithms.

Scott Levin, Reader

Todd McNutt, Reader

Sauleh Siddiqui, Advisor

Acknowledgment

I'm sincerely grateful for my advisor Dr. Sauleh Siddiqui's guidance through the course of my Ph.D. in the last three and half years. He brought me into the field of healthcare research, which is an area I'm very interested in and where I can apply machine learning and optimization techniques to solve real problems. In addition, Hopkins is a perfect place to conduct healthcare research as an engineer. I have grown significantly as a researcher under his supervision. Thank you for letting me work in this area and supporting me to explore the ideas I'm interested in.

It was my great privilege to also work with Dr. Scott Levin and Dr. Sean Barnes on the heart failure patient's readmission prediction project. I would like to thank them for their mentoring, insightful discussions, and all the effort they spent on working together on the project. This challenging work won't have been possible without their support. Thank you for leading the project, editing the draft and providing directions and suggestions during all the meetings.

I would like to thank Dr. Todd McNutt for letting me work closely with their team in the radiation oncology department at Hopkins on toxicity prediction for head and neck patients for my final year. Thank you for your mentorship and leadership. It has been my great pleasure and a wonderful experience working with the team and applying machine learning techniques to potentially improve patients' quality of life. I would like to thank Dr. Harry Quon, Dr. Ilya Shpitser, Dr. Russ Taylor, Pranav Lakshminarayanan, Peijin Han, Sierra Cheng and Xuan Hui for their help on the project. The work won't be possible without the great teamwork.

I would also like to thank my colleagues in the Mathematical Optimization for Decision Lab (MODL): Dr. Felipe Feijoo, Dr. Jimi Oke, Haralampos Avraam, Sriram Sankaranarayanan, Tom Brijs and Dr. Ying Zhang for their inspiring discussions and comments on my work. It was my great pleasure to be in the same lab with you all. My other colleagues in the department, Gary Lin, Guanbo Bian, Hwanpyo Kim, Nicolas Venkovic, Sen Lin and Xiaofan Zhang have provided me great support as companions. Thank you all for the friendship and encouragement.

I would like to thank all the staff and faculties in the Civil Engineering department. Especially, our academic coordinator Lisa Wetzelberger, and Deborah Lantry have been very supportive and helpful for my administrative requests. Special thanks to Richard D. Hickman for providing fellowship for the first year's study of my Ph.D.

Finally, I want to thank my mom and dad, my brother for their tremendous and continuous support through my Ph.D. They are the reasons that I survived the most difficult times. I will always be grateful for your love and support.

Funding

The first year (2014-2015) of my Ph.D. was funded by the Richard D. Hickman Civil Engineering Fellowship. A 2015 Johns Hopkins Discovery Award for "*Supervised Machine Learning for Clinical Decision Support for Heart Failure Readmission*" funded me from September 2015 to December 2016. My final year funding since January 2017 has been provided by the Radiation Oncology Institute (the Johns Hopkins University's grant number is 125388 and sponsor's award ID is ROI2016-912).

Dedication

To brother (Hao Jiang), Mom (Ping He), and Dad (Sizhi Jiang)

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Healthcare and energy systems are critical components of our modern society. New challenges are constantly rising within these systems. For instance, based on the World Health Organization, the United States has the highest health care cost as a percentage of its GDP (17.1%) per capita, among all nations in 2014. The same year, it was ranked by Bloomberg among the ten countries with the least efficient healthcare out of 55 countries studied. Not only is high healthcare cost a challenge, most importantly, we also want to improve the treatment effectiveness and patients' quality of life. On the other hand, for energy systems, ensuring effective renewable energy policies is crucial to mitigate climate change and reduce carbon emissions. However, reducing healthcare costs, improving patient treatment outcomes, and ensuring effective renewable energy policies are complex issues. In this dissertation, we attempted to provide our contribution to addressing a subset of these complex questions through an engineering approach, precisely, using machine learning and optimization techniques.

Machine learning and artificial intelligence is becoming more and more powerful and changing the way we live. Meanwhile, patients' data and renewable energy trade data have been accumulating with the advances in information technology. We believe using machine learning and optimization informed by these data will improve the efficiency and inform better decision making for healthcare and energy systems. Throughout the dissertation, we will demonstrate how we contributed to improving the healthcare and energy systems using machine learning and optimization techniques. We also show how we are advancing the machine learning algorithms through more efficient hyper-parameter optimization techniques.

In the next two sections, we describe a selected subset of the specific machine learning and optimization methods we mainly studied and applied for our research problems in detail. The fourth section of this chapter explains the particular research applications and contribution we made. The final section outlines the overall structure of this dissertation.

1.1 Machine Learning Methods

1.1.1 Supervised Learning

Supervised learning is a task of learning a function that maps inputs to the corresponding outputs from a given set of input-output examples [1]. Then we can use the learned function to predict the value of outputs given the inputs. The inputs and outputs are also called features and labels/prediction targets. In the field of statistics, the inputs have often been called

predictors/independent variables, and the outputs have been called responses/dependent variables [2]. The input-output pairs provided is the training data for the learning task. Depending on the variable type of the outputs, the supervised learning task is often categorized into two different types: classification and regression. If the outputs are categorical variables or qualitative variables such as color, of which the values don't have an explicit ordering, the learning task is called classification. If the outputs are quantitative variables, such as housing price, of which the values have an explicit ordering, the learning task is called regression. Many different algorithms exist for the classification and regression tasks. In the dissertation, we mainly applied and studied classification algorithms, specifically, logistic regression models, with and without ridge and lasso regularization, K-means clustering [2], stochastic gradient descent, and coordinate descent methods. Support vector machines [3] is described in detail in Chapter 5.

Let's define the input data as vectors $\boldsymbol{x}_i$ from feature space $X \in R^p$, $i \in \{1, \dots N\}$, where p is the dimension of the input vector and N is the number of training examples. Define y_i as output labels from output space $Y \in \mathbb{R}$ (for regression), or $Y \in \mathcal{Y}$ (for classification) where $\mathcal{Y}$ is the set of values for all the labels. We aim to learn a function $f(\boldsymbol{x}): X \to Y$ that can predict y. To learn the function, $f(\boldsymbol{x})$, many different learning algorithms have been developed. The learning task usually results in an optimization problem in which an empirical loss function, $L(f(\boldsymbol{x}), y)$, is used as the objective function. Various loss functions exist such as mean square error (often used for regression), hinge loss (used for support vector machines), and negative log-likelihood (used for

logistic regression). Given a specific learning algorithm and loss function, minimizing the loss on the training data using an optimization technique will yield the function $f(\boldsymbol{x})$.

1.1.1.1 Logistic Regression

The logistic regression model is a popular classifier that directly models the posterior probabilities of K classes using a logistic function, also called the sigmoid function as follows:

$$
\begin{aligned}
P(Y=1|X=\boldsymbol{x}) &= \frac{e^{(\boldsymbol{\beta}_1^T\boldsymbol{x}+\beta_{10})}}{1+\sum_{i=1}^{K-1} e^{(\boldsymbol{\beta}_i^T\boldsymbol{x}+\beta_{i0})}} \\
P(Y=2|X=\boldsymbol{x}) &= \frac{e^{(\boldsymbol{\beta}_2^T\boldsymbol{x}+\beta_{20})}}{1+\sum_{i=1}^{K-1} e^{(\boldsymbol{\beta}_i^T\boldsymbol{x}+\beta_{i0})}} \\
&\vdots \\
P(Y=K|X=\boldsymbol{x}) &= \frac{1}{1+\sum_{i=1}^{K-1} e^{(\boldsymbol{\beta}_i^T\boldsymbol{x}+\beta_{i0})}}
\end{aligned}
\tag{1.1}
$$

This logistic function formulation ensures that $P(Y=k|X) \in (0,1), k \in \{1,\dots,K\}$ and $\sum_k P(Y=k|X)=1$. $\boldsymbol{\beta}_i$ and β_{i0} are the model parameters to be learned, also called feature weights.

Logistic regression is a linear classifier as the decision function defined by $\boldsymbol{\beta}_i^T\boldsymbol{x}+\beta_{i0}$ is a linear function in the input data $\boldsymbol{x}$. As a result, the decision boundary that separates the input feature space by logistic regression is also linear. We can ignore the bias term β_{i0} to simplify the formulations by extending the parameters and input vector as follows:

$$\boldsymbol{x}_i^T \leftarrow [\boldsymbol{x}_i^T, 1], \quad \boldsymbol{\beta}^T \leftarrow [\boldsymbol{\beta}, \beta_0] \tag{1.2}$$

For binary classification tasks, i.e., $Y \in \{0,1\}$, the above formulation simply reduces to

$$P(Y = 1|X = \boldsymbol{x}) = \frac{e^{\boldsymbol{\beta}^T \boldsymbol{x}}}{1 + e^{\boldsymbol{\beta}^T \boldsymbol{x}}} \tag{1.3}$$
$$P(Y = 0|X = \boldsymbol{x}) = \frac{1}{1 + e^{\boldsymbol{\beta}^T \boldsymbol{x}}}$$

A measure of interest is the odds ratio, often used as a more intuitive estimate of the effect size of certain input feature on the outcome probabilities. For instance, to estimate the effect size of input feature x_1 on predicting $P(Y = 1|x_1, x_2, \dots, x_p)$ over $P(Y = 0|x_1, x_2, \dots, x_p)$ conditional on that the values of other features $x_2, x_3, \dots, x_p$ are fixed, the odds ratio is:

$$\frac{\frac{P(Y = 1|x_1 = 1, x_2, \dots, x_p)}{P(Y = 0|x_1 = 1, x_2, \dots, x_p)}}{\frac{P(Y = 1|x_1 = 0, x_2, \dots, x_p)}{P(Y = 0|x_1 = 0, x_2, \dots, x_p)}} = \frac{e^{(\beta_1 + \sum_{i=2}^{p} \beta_i x_i)}}{e^{(\sum_{i=2}^{p} \beta_i x_i)}} = e^{\beta_1} \tag{1.4}$$

Therefore, a positive weight β_i indicates an odds ratio larger than one and feature x_i is positively associated with predicting Y as one. A negative weight β_i indicates an odds ratio smaller than one and feature x_i is negatively associated with predicting Y as one. A zero weight β_i means the feature x_i is irrelevant in predicting Y.

The weights of the logistic regression model are typically estimated by fitting the training data by maximum likelihood estimation. The conditional likelihood of seeing $y_1, y_2, \dots, y_n$ given $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n$ is:

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{i=1}^{N} P(Y = y_i | X = \boldsymbol{x}_i; \boldsymbol{\beta}) \tag{1.5}$$

Log-likelihood is usually maximized instead for mathematical convenience.

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{N} logP(Y = y_i | X = \boldsymbol{x}_i; \boldsymbol{\beta}) \tag{1.6}$$

Next, we show how the log-likelihood function can be maximized using the Newton-Raphson algorithm with a binary classification example, which significantly simplifies the derivation.

Inserting Eq. (1.3) into (1.6) and rewriting, we obtain:

$$\begin{aligned} l(\boldsymbol{\beta}) &= \sum_{i=1}^{N} logP(Y = y_i | X = \boldsymbol{x}_i; \boldsymbol{\beta}) \\ &= \sum_{i=1}^{N} logP(Y = 1 | X = \boldsymbol{x}_i)^{\boldsymbol{y}_i} P(Y = 0 | X = \boldsymbol{x}_i)^{1-\boldsymbol{y}_i} \\ &= \sum_{i=1}^{N} \{y_i logP(Y = 1 | X = \boldsymbol{x}_i) + (1 - y_i) logP(Y = 0 | X = \boldsymbol{x}_i)\} \end{aligned}$$

$$= \sum_{i=1}^{N} \{y_i \boldsymbol{\beta}^T \boldsymbol{x_i} - \log(1 + e^{\boldsymbol{\beta}^T \boldsymbol{x_i}})\} \tag{1.7}$$

To maximize $l(\boldsymbol{\beta})$, we can set its first order condition to be 0.

$$\frac{\partial l(\boldsymbol{\beta})}{\boldsymbol{\beta}} = \sum_{i=1}^{N} \boldsymbol{x}_i [y_i - P(Y = 1 | X = \boldsymbol{x}_i)] = 0 \tag{1.8}$$

This is a system of $p + 1$ nonlinear equations and a Newton-Raphson algorithm is typically used to solve them [4].

Logistic regression is widely used because it provides well-calibrated probability estimates, which are very important for certain applications, such as risk modeling in finance. Probability estimates also make it straightforward to compute certain complex performance metrics, such as area under the receiver operating characteristic curve (AUC) [5]. As for other supervised learning algorithms, such as support vector machines that don't yield probability estimates, another algorithm needs to be used to obtain a probability measure for the outcome to compute the AUC score. As a simple linear classifier, the results of logistic regression are also easy to interpret. For instance, the *p*-values associated with the weights indicate whether certain features are statistically significantly different from zero. The sign of the weights indicates whether the outcome and the inputs are positive or negatively associated. In other words, the logistic regression model is not only used for prediction but also often used for statistical inference.

However, as an example of generalized linear models, logistic regression won't lead to valid statistical inference when the issue of multicollinearity is present, or the number of features is larger than the number of training examples. Specifically, when the number of features exceeds the number of training examples, the linear system of equations (1.8) has more variables than the number of equations. As a result, Eq. (1.8) doesn't have unique solutions and the solutions for the weights have high variance (sensitive to slight changes in input data). The issue makes it unreliable for statistical inference. Besides, the high variance may lead to poor generalization performance on unseen test data [6]. Next, we introduce two regularization techniques that are often used to deal with this issue.

1.1.1.2 Ridge and Lasso Regularization

Ridge and lasso regularization are two general regularization techniques. They are often used to control model complexity and prevent overfitting for supervised learning models. They are widely used with many different supervised learning algorithms such as generalized linear models, support vector machines, and neural networks. Ridge regularization adds the l^2-norm of the weights as a penalty term to the original loss function, while lasso regularization adds the l^1-norm of the weights as a penalty term [2][6]. Next, we use logistic regression as an example to describe ridge and lasso regularization.

The loss function, negative log-likelihood of logistic regression is provided in Eq. (1.7). For ridge regularization, after adding the l^2-norm of the weights to the loss function, we obtain the new loss function:

$$L(f(\boldsymbol{x}), y) = l(\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|^{\mathbf{2}} = l(\boldsymbol{\beta}) + \lambda \sum_{i=1}^{p} \beta_i^2 \tag{1.9}$$

For lasso regularization, we add the l^1-norm of the weights to the loss function and obtain the new loss function:

$$L(f(\boldsymbol{x}), y) = l(\boldsymbol{\beta}) + \lambda \sum_{i=1}^{p} |\beta_i| \tag{1.10}$$

where λ is the regularization hyper-parameter that controls the extent of regularization. As a hyper-parameter, it has to be exogenously set before we train the logistic regression model. In practice, the hyper-parameter is often tuned by cross-validation to prevent overfitting.

By adding a penalty term of the weights into the objective functions, ridge and lasso increase the bias of the estimation but reduce the variance of the estimation as the penalty term shrinks the value of the weights. How much variance ridge and lasso reduces is controlled by the value of the regularization hyper-parameter λ. A large value of λ will lead to very small values of the weights β_i.

The advantage of ridge and lasso is that they both work well for high dimensional feature space data, where the number of features exceeds the number of training examples. By reducing

the variance of the estimation, they can prevent overfitting and improve the model's generalizability on unseen test data. The difference between these two methods is that lasso induces sparsity into the solution due to l^1-norm regularization but ridge doesn't. As a result, the lasso is often used for feature selection, i.e., selecting a subset of predictive features from the original feature set. Ridge regularization uses a smooth l^2-norm regularization. Thus, it doesn't induce a sparse solution, rather shrinks the weights of nonimportant features close to zero and penalizes large weights. Another difference between ridge and lasso is their usage for statistical inference given correlated input features. According to Friedman et al. (2010)'s study [7], ridge regularization shrinks the weights of correlated features towards each other (it will assign equal weights to k identical input features). Ridge works well for the case when there are many input features and all of them have a non-zero effect on the output. On the other hand, lasso tends to select one of the correlated features and set the weights of the other correlated features as zero [7]. These characteristics are further explored in Chapter 3. A coordinate descent method has also been successfully applied to solve generalized linear models with ridge and lasso regularizations [2][6][7].

1.1.2 Unsupervised Learning

Unsupervised learning is a task in which an algorithm or agent learns the patterns in input data while no explicit output is provided [1]. As the most common unsupervised learning task, clustering identifies groups or clusters that potentially exist in the input examples. The input

examples identified to be within the same cluster are close to each other given certain distance or dissimilarity measure. In this thesis, the K-means clustering method was used in Chapter 2 and we will describe it in detail here.

The K-means clustering algorithm is an iterative method that uses the squared Euclidean distance as its dissimilarity measure [2]. The squared Euclidean distance between input example x_i and x_j is:

$$D(x_i, x_j) = \|x_i - x_j\|^2 = \sum_{k=1}^{p} (x_{ik} - x_{jk})^2 \tag{1.11}$$

The within-cluster sum of squares (WCSS) measure the total distance between all the input examples within a cluster and the center or mean vector of that cluster. Let's define μ_i as the mean vector of the cluster S_i and there are k clusters. Then the total WCSS for the input data is:

$$WCSS = \sum_{i=1}^{k} \sum_{x_j \in S_i} \|x_j - u_i\|^2 \tag{1.12}$$

K-means iteratively minimize this WCSS as follows:

K-means Algorithm

1. Choose the number of clusters k. Initialize the algorithm by randomly assign a label from the k cluster labels to each of the input examples.

2. Compute the cluster center/mean vector for each cluster given the cluster assignment,

   $\mu_i^t = \frac{1}{|S_i|} \sum_{x_i \in S_i} x_i$

3. Assign each input example to the nearest cluster, i.e.,

   $\|x_i - \mu_p^t\|^2 \leq \|x_i - \mu_j^t\|^2 \; \forall j: 1 \leq j \leq k.$

   where p is the new cluster label for input example x_i.

4. Repeat step 2-3 until the cluster assignment doesn't change.

As this is a heuristic and the optimal solution is not guaranteed, often the algorithm is repeated multiple times with different random initializations in step 1. The results with the lowest WCSS is picked as the final cluster assignment result. In practice, the elbow method, or formally, gap statistic is typically used to estimate the optimal number of clusters [8].

In Chapter 2, we used this algorithm to perform clustering on the time series data of predicted daily readmission risk. As a result, we detected different dynamic risk readmission patterns from admission to discharge.

1.2 Optimization Methods

1.2.1 Stochastic Gradient Descent

In this big data era, the increase of data size often exceeds the increase of processor speed. Fast and simple optimization methods that work well for large-scale data problems have been a popular research topic for the last decade. First-order gradient methods have seen great success in solving large-scale problems. Among them, the stochastic gradient descent (SGD) is the most popular and an asymptotically efficient method, especially as the technique predominantly used for solving deep neural networks [9].

Let's denote the loss function for our machine learning task as $L(f(x;\beta),y)$, where the function $f(x;\beta)$ is what we are learning to make predictions, β is the model parameters, x and y are the input and output of the data examples. Let's denote the distribution of data examples as $P(X)$ which is often unknown in practice. To learn function $f(x;\beta)$, we want to minimize the expected loss over the distribution $P(X)$, i.e., $E(f)=\int L(f(x_i;\beta),y_i)dP(X)$. As $P(X)$ is unknown, the empirical risk minimization is often performed [9]. The empirical risk is $E_n(f)=\frac{1}{n}\sum_1^n L(f(x_i;\beta),y_i)$, where n is the number of data examples.

Gradient descent has been proposed to minimize the empirical risk [10]. The gradient of $E_n(f)$ with respect to the weight parameter β is $\nabla_\beta E_n(f)=\frac{1}{n}\sum_1^n \nabla_\beta L(f(x_i;\beta),y_i)$. Gradient

descent is an iterative numerical optimization algorithm which updates the weight parameter β in each iteration using:

$$\beta^{t+1} = \beta^t - \alpha \frac{1}{n} \sum_{1}^{n} \nabla_\beta L(f(x_i; \beta), y_i) \tag{1.13}$$

where α is the step size chosen to update the weight parameter.

As an even simpler method, instead of computing the full gradient of $\nabla_\beta E_n(f)$ using all the data examples, SGD estimates the gradient $\nabla_\beta E_n(f)$ by randomly picking a single data example x_t. Then each iteration of updating the weight parameter becomes:

$$\beta^{t+1} = \beta^t - \alpha_t \nabla_\beta L(f(x_t; \beta), y_t) \tag{1.14}$$

It's shown $\nabla_\beta L(f(x_t; \beta), y_t)$ is an unbiased noisy gradient estimate of $\nabla_\beta E_n(f)$ and in expectation, parameter update Eq. (1.14) will be the same as Eq. (1.13). Under conditions: $\sum_t \alpha_t^2 < \infty$ and $\sum_t \alpha_t = \infty$, SGD will converge to the optimum solution [9].

For large datasets, estimating one full gradient is computationally expensive, while the gradient update in SGD is much cheaper. The advantage of SGD is that it's simple and efficient. The disadvantages are that it requires tuning of some hyper-parameters, e.g., the step size/learning rate α_t, and it's sensitive to input feature scales. We used SGD to optimize the hyper-parameter of support vector machines, which is detailed in Chapter 5.

1.2.2 Coordinate Descent

Like gradient descent, coordinate descent (CD) [11] is also an iterative numerical optimization method. It has also gained recent popularity due to its successful application to solve machine learning problems, for instance, optimizing large-scale support vector machine problems [12]. Unlike gradient descent, in each iteration, CD chooses to optimize only one parameter variable while fixing the values of all the other parameter variables. Various mechanisms exist to cycle through all the parameter updates. During each iteration, CD is a solving a subproblem and often much easier than solving the full problem, for instance, when you can simply analytically derive the optimal solution for the subproblem but not for the full problem.

Using the previous empirical risk minimization problem as an example, the general steps for CD are [11]:

Coordinate Descent Algorithm

1. Initialize k to be 0, and $\beta^0 \in \mathbb{R}^p$
2. Repeat:
   - 2.1 Choose an index $i_k \in \{1,2,\dots,p\}$, which is the index for one parameter variable in the parameter vector β.
   - 2.2 $\beta^{k+1} \leftarrow \beta^k - \alpha_k \left[\nabla_\beta E_n\left(f_{\beta^k}\right)\right]_{i_k} e_{i_k}$ for some $\alpha_k > 0$ and e_{i_k} is a vector whose i_k's element is one and all other elements are zero.
   - 2.3 $k \leftarrow k + 1$
3. Until stopping criteria satisfied.

The CD method was studied and successfully applied in Chapter 5.

1.3 Applications and Contribution

In this section, we describe our use of the above machine learning and optimization techniques. Also, we highlight the challenges and complexity of each project and explain how we applied the machine learning and optimization techniques to deal with these challenges. Finally, we

summarize the contribution and importance of our work in each project, as well as the overall contribution of the dissertation.

1.3.1 Machine Learning in Healthcare

Machine learning techniques have advanced rapidly and been successfully applied to solve problems in various fields such as natural language processing, computer vision, and robotics. Recently, with the increasingly available healthcare data, researchers have been actively applying machine learning techniques to solve problems in healthcare, such as disease diagnosis and treatment outcome risk prediction.

Unlike the application of machine learning in other fields, machine learning application in healthcare faces several challenges and also opportunities. First, patients' life and safety are so important that any prediction model built using machine learning techniques needs to be very accurate to be practically used in clinical settings. The users of the prediction model have to be able to control the false negative rate and false positive rate. For instance, for cancer diagnosis, if a cancer patient was falsely diagnosed as not having cancer by the prediction model, the patient could die or miss the early treatment. On the other hand, if a non-cancer patient was falsely diagnosed as having cancer by the prediction model, the patient could receive unnecessary treatment that could cause side effects and higher healthcare costs. In healthcare settings, a false negative diagnosis is much worse than a false positive diagnosis made by the prediction model.

Therefore, it's extremely important for the physicians to be able to control for the false negative rate.

Second, the high complexity of human disease and biological mechanism make applying machine learning models very challenging. Machine learning may be able to provide accurate predictions or discover new knowledge in healthcare, but physicians' expertise will always be required to make actual decisions to treat the patients. Therefore, how to bridge the gap between machine learning models physicians' decision making is a key step to successfully leveraging machine learning techniques in healthcare. For instance, assume we have obtained an accurate readmission risk prediction for a patient, the physicians still need to decide what interventions to provide to effectively reduce the readmission risk for that patient. A prediction model that also provides important potential driving factors that affect the readmission risk will be informative for physicians' decision-making, while a prediction model itself is informative enough in this case.

Third, each patient is unique and has different characteristics from other patients. The same drug may be effective for one patient but not for another patient. Therefore, precision medicine was proposed to provide customized medical treatment for each patient. To achieve the goal of precision medicine, machine learning techniques are promising as they can take each patient's characteristics; such as genomic data, clinical history, and diagnosis information; into account and provide customized prediction or suggestions for treatment. This is also the advantage of machine

learning models over traditional statistical analysis, which typically draws conclusions on the population level.

In conclusion, machine learning techniques are promising and have great potential to advance the field of healthcare but face critical challenges that need to be overcome to prove its practical value.

1.3.2 Knowledge Discovery and Prediction

Machine learning and optimization methods can be used both for knowledge discovery and prediction. In many applications in industry, such as demand forecasting, loan default prediction, and advertisement click-through rate prediction, an accurate prediction is most important. The user may care less about discovering new knowledge from the database. However, that's often not the case in the context of medicine.

In medicine, the physicians not only need an accurate prediction model for tasks such as toxicity prediction or disease diagnosis but are also interested in discovering new knowledge about the toxicity or disease. The physicians want to investigate the clinical and administrative factors that are causing toxicity and diseases. Often knowledge discovery is more important than obtaining an accurate prediction model. If we know the important factors which are causing certain outcomes, the physicians can then provide effective targeted interventions to improve patients' treatment outcome. Merely knowing an accurate prediction for a patient is not good enough since the physicians still need to make further decisions when the predicted outcome is bad. In other

words, knowledge discovery may be more valuable than prediction regarding helping physicians with their decision making.

Knowledge discovery is a task where the users want to discover new knowledge or pattern from the database. Various machine learning models have been used for knowledge discovery. For instance, using logistic regression, lasso regularization, or random forest models to find the most critical features and perform feature selection; or using clustering algorithms to find potential patterns and clusters in the dataset.

Unlike knowledge discovery, the goal of prediction is to predict an outcome Y given input data X using a function $f(X)$. To achieve good prediction performance, often complex black-box models such as deep neural networks are used, which are difficult to interpret. On the other hand, knowledge discovery often utilizes more interpretable models rather than black-box machine learning models, as an accurate prediction is not the primary goal.

For our heart failure readmission risk prediction model, we focus both on accurate prediction and knowledge discovery regarding the driving factors for readmission. For the radiation oncology project, we mainly focus on knowledge discovery regarding the investigating the influence of spatial radiation dose pattern on patients' xerostomia outcome.

1.3.2.1 Readmission Risk Prediction

Heart failure is a leading cause of hospitalization with a high 30-day readmission rate of about 20%. A high hospital readmission rate indicates poor patient treatment outcomes and increased

healthcare cost. The goal of this project is to build a dynamic daily prediction model for 30-day readmission risk for heart failure patients. The prediction model can help hospitals target patients that have high readmission risks with interventions to reduce readmission rate. The challenge is that the readmission risk (i.e., the outcome label), if the patients were discharged earlier, is unknown. Another challenge is that predicting readmission risk itself is a hard problem and existing models usually have poor prediction performance. To deal with these challenges, we came up with a two-stage modeling approach in which we combined classification and regression techniques. More specifically, at the first stage, we trained a classification model to estimate the counterfactual daily readmission risk; at the second stage, we fitted a beta regression model using patient-day data and the estimated readmission risk outcome from the first stage. Finally, we obtained a dynamic daily readmission risk prediction model that can be used in practice, which is the trained beta regression model at the second stage. Further, we found four different risk readmission groups with different dynamic risk trajectories using clustering techniques. We also identified predictors that are most associated with the changing readmission risk. Our contribution is that due to unavailable true daily readmission risk data, we combined a classification model with a regression model to dynamically predict readmission risk and clustered different readmission risk groups.

1.3.2.2 Knowledge Discovery in Radiation Oncology

Head and neck cancer patients often receive radiation therapy to kill cancer cells. However, the head and neck contain several organs and tissues that should not be overexposed to this radiation. Radiation induces side effects such as xerostomia, which is also called dry mouth. The goal of this work is to understand how radiation dose affects xerostomia for head and neck cancer patients. There are several main challenges for this work. First, there is not enough variation of radiation dose treatment in different subvolumes of organs across the patients, which makes it hard for the machine learning algorithms to detect the most sensitive regions. Second, the dose effect on xerostomia in different regions may have spatial dependency and radiation dose data alone is unable to identify the sensitivity of dose effect. To deal with those challenges, we performed a voxel-based radiation dose analysis by combining a radio-morphology model and machine learning algorithms. Specifically, we tried various machine learning algorithms (logistic regression with ridge, lasso regularization, random forest) and found out that ridge logistic regression is well suited to identify the influential subvolume regions given the high dimensional and spatially highly correlated radiation dose data. However, to further establish causal effect between dose and xerostomia, causal inference and clinical trials need to be conducted. For this work, our contribution is that, given limited patients' radiation treatment data, we combined a radio-morphology model with regularized supervised learning algorithms to learn dose effect on xerostomia for head and neck cancer patients. Notably, we found that the dose level in the low

dose bath region, i.e., the superior portion of ipsilateral and contralateral parotid glands, to be most influential on acute xerostomia and xerostomia recovery.

1.3.3 Renewable Energy Policy Analysis

Traditionally, wood chip was used to produce paper, but it has also been used as biofuel. The US is considering including wood chip in its renewable fuel standard, which is a policy to reduce overall carbon emissions. But the US has been a major exporter of wood chip worldwide, and its exports of wood chip could decline if this resource is increasingly used for domestic electricity generation and biofuel production due to this policy. The research question we are investigating here is how local renewable energy policies in the US can affect global wood chip trade and carbon emissions. Through global trade of renewable bioenergy products, local renewable energy policies can impact the bioenergy production in other world regions. We need to estimate the global impacts of the local energy policies before we implement them. The challenge of this work is that the wood chip trade data is very limited. We were able to automatically calibrate our wood chip trade model under this data-limited setting using a primal-dual optimization technique. For this project, our contribution is that we build the first global wood chip trade model in the literature by combining a fundamental economic equilibrium model with top-down optimization methods to deal with limited data.

1.3.4 Hyper-parameter Optimization

Hyper-parameter optimization is crucial to obtain a trained machine learning model that has good prediction performance. However, in practice, the methods used for hyper-parameter optimization are often empirical or ad hoc search methods such as grid search and random search. The challenge is that it's computationally expensive to search the hyper-parameter space and more efficient methods are needed. We adopted a gradient-based approach and combined stochastic gradient descent and dual coordinate descent to optimize the hyper-parameter for Support vector machines. The contribution of this work is that we used fundamental gradient-based optimization methods for hyper-parameter optimization for a machine learning algorithm that enabled hyper-parameter tuning to evolve away from empirical ad hoc approaches.

In summary, the overall contribution of this dissertation is that we integrated bottom-up, fundamentals-based models with top-down machine learning and optimization models in data-limited settings with applications in healthcare and energy systems. We also advanced the method for hyper-parameter tuning for machine learning algorithms using optimization techniques.

1.4 Outline

Chapter 2 describes our work in building dynamic 30-day readmission risk prediction model for heart failure patients. It applies both supervised and unsupervised machine learning methods as well as feature selection techniques. This work was mainly advised by Dr. Scott Levin, Dr. Sean

Barnes, and Dr. Sauleh Siddiqui. This chapter was written for an academic journal and is planned to be submitted this year. Chapter 3 describes our work in the radiation oncology department. We demonstrated how to use machine learning methods to perform knowledge discovery on the relationship between radiation dose on radiotherapy side effects, particularly xerostomia, for head and neck cancer patients. The first half of the results of this chapter have been submitted for journal publication and are under review. This work was advised by Dr. Todd McNutt, Dr. Harry Quon, Dr. Sauleh Siddiqui, Dr. Ilya Shpitser, Dr. Russell Taylor, and completed together with Pranav Lakshminarayanan, Xuan Hui, Peijin Han, Sierra Cheng, who are our colleagues at the radiation oncology department at Johns Hopkins University. Chapter 4 describes our work on applying equilibrium modeling and optimization techniques for analysis on renewable policies of wood chip. This work was advised by Dr. Stephanie Searle and Dr. Sauleh Siddiqui and has been published as an academic journal article. Chapter 5 shows how we applied optimization techniques to improve the efficiency of building a support vector machine classifier, which is a popular supervised machine learning algorithm. The research was mainly advised by Dr. Sauleh Siddiqui. We plan to submit this work to a journal in the field of operations research.

Finally, we summarize the overall findings and conclusions of this dissertation, and also provide future directions encouraged by this dissertation in Chapter 6.

Chapter 2

Heart Failure Patient Readmission Prediction

2.1 Introduction

Patients hospitalized with heart failure suffer the highest rates of 30-day readmission among any clinically-defined patient populations in the United States (US) [13]. National efforts to prevent avoidable hospitalizations have led to the adoption of 30-day readmission as a publicly reported performance measure linked to Medicare patient reimbursement [14]. This has motivated much investigation into the predictability of 30-day readmissions in-hospital to guide targeted interventions that could reduce risk. Clinical and administrative data available in hospital electronic health records (EHR) and clinical registries have been the primary data sources for these evaluations. Prior studies have applied traditional statistical- and machine-learning-based methods for predicting 30-day readmission with varied success for heart failure patient cohorts; predictive

performance measured as the area under the receiver operating characteristic curve (AUC) has ranged from 0.55 to 0.76 [15][16][17][18][19][20].

In this study, a novel two-stage modeling approach to estimating patients' readmission risk dynamically was created for a cohort of heart failure patients presenting to a community hospital. The study objective was to develop a predictive model of 30-day readmission that functions in real-time over the course of a patient's hospitalization. We hypothesize that quantifying readmission risk trends has potential to illuminate the effects of clinical measures and interventions on readmission likelihood at discharge. This analysis further enabled the identification of heart failure patient groups with fundamental trajectories in readmission risk over their hospital encounter.

2.2 Methods

2.2.1 Setting and Data Source

We conducted a retrospective cohort study of patient encounters with a primary diagnosis of heart failure between September 1, 2013, and August 31, 2015, from a community hospital in Columbia, Maryland. We identified heart failure patients using ICD9 codes: 428.x, 402.01, 402.11, 402.91, 404.01, 404.03, 404.11, 404.13, 404.91, and 404.93 [21]. Encounters that resulted in mortality, transfer to other acute care hospitals, or hospice care were excluded. The final cohort consisted of

534 unique encounters for 478 patients totaling 2750 patient-days. Data from the hospital EHR available as part of routine patient care was mined to establish our outcome and predictor variables.

We employed a Centers for Medicare and Medicaid Services (CMS) [14] definition for 30-day readmission as our primary outcome measure. This was defined as readmission to the same hospital within 30 days of discharge of the index hospitalization for all causes (i.e., not necessarily related to heart failure) [14].

Hypothesized predictors were derived from heart failure and readmission risk factors used in previous studies as summarized in Table **2.1** [17][22][23][24][25][26][27]. Predictor data may be conceptually grouped into static (unchanged over the hospital encounter) or dynamic (fluctuating over the hospital encounter). Static predictors comprised demographics (age, gender, race), socio-economic status (insurance, marital status, zip code), healthcare utilization (discharge disposition, number of visits in the last six months), emergency department (ED) chief complaint clinically categorized [28][29], Charlson co-morbidity index at admission computed using active problems on patients' problem list (medical history) [23][24], and admission diagnoses.

Dynamic predictors comprised the elapsed length-of-stay at prediction time, vital signs (systolic blood pressure, diastolic blood pressure, temperature, respiratory rate, pulse, peripheral capillary oxygen) and lab results (Alanine Transaminase (ALT), Aspartate Transaminase (AST), Blood Urea Nitrogen (BUN), Creatinine, Hemoglobin, Potassium, Sodium, proBNP, Troponin T). Longitudinal vital signs and laboratory result data accumulate over a patients' encounter. Time

series characteristics of these data were engineered as predictors. Each vital sign and laboratory measure ($x_1, x_2, \ldots x_t$) was transformed into the following predictors:

1. Number of measurements: t
2. Average value: $\bar{x} = \frac{1}{t}\sum_{i=1}^{t} x_i$
3. Standard deviation: $\sqrt{\frac{\sum_1^n (x_i - \bar{x})^2}{t-1}}$
4. Minimum: $\min(x_1, x_2, \ldots, x_t)$
5. Maximum: $\max(x_1, x_2, \ldots, x_t)$
6. Normalized index of the minimum: $\frac{index\ of\ \min(x_1, x_2, \ldots, x_t)}{t}$
7. Normalized index of the maximum: $\frac{index\ of\ \max(x_1, x_2, \ldots, x_t)}{t}$
8. Average of last three measurements: $\bar{x} = \frac{1}{3}\sum_{i=t-2}^{t} x_i$
9. Average of first-order difference: $\frac{1}{t-1}\sum_{i=2}^{t} |x_i - x_{i-1}|$

Normalized index predictors were designed to capture the relative point during hospitalization these extreme measurements occurred [30]. For example, if ten temperature

measurements have been taken over a patients' stay at prediction time t and the 6^{th} measurement is where the minimum occurred - the value of the normalized index of minimal temperature is 0.6.

Missing values were observed for laboratory result data. For these cases, the patient may not have had certain tests conducted, or the test was conducted, but the results were not recorded into the EHR system. For the first case, we have, for instance, 70 out of 534 encounters did not have pro-B-type Natriuretic peptide test data, 55 encounters did not have Aspartate Transaminase test data, and 37 encounters did not have Troponin-T test data. Features with more than 30% data missing were excluded. For features with less than 30% missing data, we imputed the missing values using the mean value for the predictors across all the 534 encounters. For the second case, we hypothesized that the number of tests conducted might reflect the severity of a patient's condition. Fewer tests may indicate a better patient condition. Therefore, we created predictors for the number of laboratory tests conducted.

2.2.2 Daily Prediction Model

The model was structured to yield a prediction of 30-day readmission risk each day (6 a.m.) in the patients' encounter based on data available at that time-point in the EHR. Predictions were chronologically made at the time of hospital admission (first prediction), then 6 a.m. each subsequent day, and last at hospital discharge (last prediction). Thus, predictions made on day 1 (6 a.m. day after admission) would have access to less predictor information than predictions made for that same patient on day 3.

To obtain a daily prediction model, we used a two-stage modeling approach. In the first stage, we trained a classification model on the 534 encounter samples with their 30-day readmission labels as prediction target. Then, we used the trained classifier to estimate the daily readmission probability using patients-day data. Patients-day data consists of patients' information accumulating until a certain day. In the second stage, we fit a beta regression model using patients-day data as predictors and the estimated daily readmission probability as prediction target. There are two main reasons for using this two-stage approach. First, to ensure that training data and future daily prediction data are from the same distribution. Second, the counterfactual 30-day readmission risk if the patients were discharged earlier is unknown, and we used the first stage model to estimate it.

Multiple classification algorithms (e.g., logistic regression, random forest, AdaBoost) were used to train the first stage model with logistic regression with backward stepwise selection yielding the best performance. The performance was evaluated out-of-sample through 5-fold cross-validation (80% training and 20% test).

2.2.3 Patient Trajectory Cluster Analysis

A patient's daily 30-day readmission risk may change over the course of their hospitalization due to increased collection of information (i.e., more data available) and the evolution of their condition as therapies are administered. We applied K-means clustering (unsupervised) to determine readmission risk trajectories over the first five days of patients' hospital stay [31]. A

total of 75.0% of all laboratory tests, 60.5% of all procedures, and 85.9% of all medication orders occurred before the end of day 5 for our study cohort. Thus, the majority of variation in readmission probability occurred before this time. The unsupervised clustering approach was designed to learn potential trajectories (i.e., trends) in readmission risk that may naturally distinguish patient groups. Predictors that discern these different trajectories were then detected using Kruskal Wallis hypothesis testing.

The specific steps used to train the predictive models, learn patient groups by trajectory, and identify predictors that discern these patient groups are summarized as follows:

1. Train logistic regression model on the full data set using the selected set of predictors
2. Estimate the probability of 30-day readmission for each patient-day using the trained classification model
3. Fit a beta regression model using the patients-day data with same selected set predictors in step 1 as predictors and estimated readmission probability from step 2 as prediction target
4. Assemble the time series of daily readmission probabilities predicted by the beta regression model for each patient encounter
5. Apply the K-means algorithm to cluster encounters into groups based on their respective time series trajectories of daily readmission probabilities

6. Apply Kruskal-Wallis test to identify discriminative predictors associated with the different clusters

Analyses were conducted using Python (version 2.7), the scikit-learn (version 0.18.1) and R (version 3.4.1), betareg (version 3.1-0).

2.2.4 Partial Dependence of Predictors

To investigate how individual predictors affect the predicted readmission risk in the fitted beta regression model, we created partial dependence plots. Partial dependence plots were proposed by Friedman (2001) to visualize the dependence of the prediction target $\widehat{F(\boldsymbol{x})}$ on the predictors $\boldsymbol{x}$, i.e., how $\widehat{F(\boldsymbol{x})}$ changes as a function of $\boldsymbol{x}$ [32]. $\widehat{F(\boldsymbol{x})}$ is the prediction model, in our case, the fitted beta regression model, and $\boldsymbol{x}$ is the set of predictors.

Assume we want to investigate the dependence of $\widehat{F(\boldsymbol{x}_s)}$ on a subset of predictors $\boldsymbol{x}_s$, which we call partial dependence. Denote the complement set of predictors as $\boldsymbol{x}_c$. We have $\boldsymbol{x}_c \cup \boldsymbol{x}_s = \boldsymbol{x}$ and $\boldsymbol{x}_c \cap \boldsymbol{x}_s = \emptyset$. To compute $\widehat{F(\boldsymbol{x}_s)}$, we marginalize out the effect of $\boldsymbol{x}_c$ on $\widehat{F(\boldsymbol{x})}$. We have:

$$\widehat{F(\boldsymbol{x}_s)} = E_{\boldsymbol{x}_C}\left[\widehat{F(\boldsymbol{x})}\right] = \frac{1}{N}\sum_{i=1}^{N} F(\widehat{\boldsymbol{x}_s, \boldsymbol{x}_{i,C}})$$

where N is the sample size of training data. In other words, $\widehat{F(\boldsymbol{x}_s)}$ is the averaged predicted value across all the training data while substituting predictor of interest $\boldsymbol{x}_s$ with a certain value for all the training samples.

2.3 Results

2.3.1 Setting and Data Source

Characteristics of the patient cohort stratified by the 30-day readmission outcome may be seen in Table 2.1. The logistic regression prediction model yielded an out-of-sample AUC of 0.73 (±0.08). Backward stepwise selection identified an optimal subset of 20 predictors (57 total). The pseudo-R-squared value (squared correlation coefficient between outcomes and predicted values) for the beta regression is 0.88.

Table 2.1 Descriptive summary of encounter data, stratified by 30-day readmission outcomes. Summary statistics for continuous variables (indicated by '*') include the mean, median, and interquartile range. Some entries were left blank (marked as '/') for the following predictors: predictors containing patient private information, complex predictors that are not straightforward to summarize in a single table such as time series predictors and the categorical features of which each patient has multiple values.

Predictor	**Descriptive statistics**		
	Not readmitted	**Readmitted**	***p*-values**

	(n = 427)	**(n = 107)**	
Static predictors			
Age	74.5, 77 (66-85)	75.2, 78 (67-87)	0.37
Gender: Female	52.2%	59.8%	0.19
Marital status			0.81
Married	39.3%	41.1%	
Single	15.9%	17.8%	
Widowed	34.4%	33.6%	
Race			0.96
White or Caucasian	60.4%	60.7%	
Black or African American	28.3%	29.0%	
Asian	8.4%	8.4%	
Insurance			0.46
Medicare	70.2%	74.8%	
Commercial	21.5%	19.6%	
Medicaid	3.0%	3.7%	
Other	5.2%	1.9%	
Discharge disposition			0.35
Home or Self Care	64.6%	60.7%	
Skilled Nursing Facility	14.5%	12.1%	
Home-Health Care Service	11.5%	15%	

Rehabilitation Facility	4.0%	5.6%	
Short Term Hospital	1.9%	1.9%	
Nursing Facility	1.2%	4.7%	
Number of prior visits to hospital*	1.4, 1 (0-2)	2.2, 1 (0-3)	0.002
Chief complaints			0.91
Shortness of breath	66%	59.8%	
Chest pain	6.1%	6.5%	
Edema	4%	5.6%	
Weakness	3.5%	2.8%	
Lower respiratory tract infection	2.8%	1.9%	
Abdominal pain	2.1%	2.8%	
General	1.6%	0.9%	
Altered mental status	1.4%	0.9%	
Genitourinary	1.2%	1.9%	
Blunt trauma	1.2%	2.8%	
ZIP code	/	/	
Diagnoses	/	/	
Diagnoses history	/	/	
Dynamic predictors			
Elapsed length of stay*	4.8, 3.7 (2.2-5.8)	6.4, 4.2 (2.6-7.3)	0.02
Laboratory test			

Average ALT (Units/L) per patient*	32.0, 20.0 (13.3-30.0)	28.0, 18.0 (12.0-28.6)	0.24
Average AST (Units/L) per patient*:	32.7, 24.0 (18.0-35.0)	32.3, 21.8 (16.9-30.0)	0.19
Average BUN (mg/dL) per patient*	28.5, 25.0 (18.3-35.3)	34.2, 30.8 (20.0-44.7)	6.05e-3
Average Creatinine (mg/dL) per patient*	1.6, 1.2 (1.0-1.6)	1.9, 1.4 (1.0-2.0)	3.79e-3
Average Hemoglobin (mg/dL) per patient*	11.4, 11.2 (9.9-12.7)	10.8, 10.4 (9.1-12.3)	2.55e-3
Average Potassium (mmol/L) per patient*	4.2, 4.1 (3.9-4.4)	4.2, 4.1 (3.9-4.4)	0.83
Average Sodium (mmol/L) per patient*	139.0, 139.0 (137.0-141.5)	137.7, 137.8 (134.8-140.9)	3.16e-3
Average proBNP (pg/mL) per patient*	8182.5, 4283.0 (2137.0-8656.0)	8683.3, 4262.0 (2013.0-9545.0)	0.92
Average Troponin T (pg/mL) per patient*	37.1, 10.0 (10.0-30.0)	56.1, 14.2 (10.0-52.7)	0.14
Vital signs			
Systolic blood pressure	/	/	
Diastolic blood pressure	/	/	
Temperature	/	/	
Respiratory rate	/	/	
Pulse	/	/	
Peripheral capillary oxygen	/	/	

Weight	/	/
Past medical history	/	/
Medication order	/	/
Procedure	/	/

2.3.2 Prediction Model

2.3.2.1 First Stage Model

As we mentioned in the previous methods section, we used a logistic regression model at the first stage. Table 2.2 summarizes the significant variables from the logistic regression prediction model at the first stage. Most of the significant predictors in Table 2.2 are clinical predictors such as laboratory test results and vital signs, specifically related to potassium, sodium, hemoglobin and diastolic blood pressure measurements. The number of medication orders for the digitalis glycosides pharmacy class, number of measurements of peripheral capillary oxygen saturation (SPO2), and number of hemodialyses performed are negatively associated with readmission likelihood. The number of measurements of sodium and number of mechanical ventilation procedures are both positively associated with readmission likelihood. Discharge disposition is a significant administrative predictor. The patients who were discharged to nursing facility compared to the non-nursing facility have a much higher readmission probability. This is probably

because patients who were sicker are more likely to be sent to the nursing facilities. Another reason maybe patients cared by nurses were readmitted to hospitals more promptly.

Table 2.2 Summary of significant predictors (*p*-values less than 0.05) for the logistic regression model with backward stepwise feature selection at discharge. The results are from the logistic regression model fitted on the entire dataset (534 encounters).

Predictor	**Coefficient**	**Odds ratio**	**95% CI**	***p*-values**
Normalized index of minimal Potassium	-2.15	0.12	-3.27, -1.03	<0.001
Number of measurements of SPO2	-0.07	0.93	-0.11, -0.03	0.001
First minus last value of Hemoglobin	0.56	1.75	0.20, 0.92	0.002
Average Sodium	-0.47	0.62	-0.78, -0.16	0.003
Number of medication order: Digitalis Glycosides	-0.67	0.51	-1.11, -0.23	0.003
First minus last value of BUN	-0.04	0.96	-0.07, -0.01	0.004
Discharge Disposition: Nursing Facility	2.36	10.59	0.62, 4.10	0.008
Number of procedures: Mechanical Ventilation	0.08	1.08	0.02, 0.14	0.008
First minus last value of Diastolic Blood Pressure	-0.02	0.98	-0.04, -0.01	0.008
Minimal Sodium	0.35	1.60	0.07, 0.64	0.01
Number of measurements of Sodium	0.22	1.25	0.05, 0.40	0.01

Normalized index of maximal Hemoglobin	1.63	5.14	0.30, 2.98	0.02
Number of measurements of Diastolic Blood Pressure	0.05	1.05	0.01, 0.09	0.02
Normalized index of minimal Sodium	-1.21	0.30	-2.25, -0.18	0.02
Average value of last 3 Diastolic Blood Pressure	-0.05	0.96	-0.09, -0.004	0.03
Normalized index of minimal Respiratory Rate	1.12	3.07	0.10, 2.15	0.03
ZIP code: 210XX	0.77	2.17	0.05, 1.50	0.04
Number of procedure: Hemodialysis	-0.41	0.67	-0.80, -0.01	0.04

2.3.2.2 Second Stage Model

After estimating the counterfactual daily readmission risk, we used beta regression as the second stage model. Beta regression models are often used to model variables that the values are in the range of $(0,1)$ in practice [33]. There are two main assumptions behind beta regression. First, the dependent variable follows a beta distribution. Second, its mean value can be fitted using a linear predictor consists of the dependent variables/predictors.

We used beta regression for the following two reasons. First, the values of the predicted daily readmission risk from the first stage model are in the interval of $(0,1)$. Second, we found the predicted readmission risk follows a beta distribution using a statistical goodness-of-fit test. Figure

2.1 shows that the predicted readmission risk can be approximately represented using a beta distribution. Q-Q plot is the quantile-quantile plot [34]. It helps diagnose whether two datasets follow the same distribution. In our case, its y-axis represents the values of different quantiles in the empirical data (estimated readmission risk), and its x-axis is the value of the corresponding quantile in the fitted theoretical beta distribution. P-P plot is the probability-probability plot [35]. It helps diagnose if an empirical distribution follows a certain theoretical distribution. The difference between P-P plot and Q-Q plot is that, instead of using quantiles, P-P plot plots the values of the cumulative distribution functions of the empirical and theoretical distributions. The y-axis and x-axis of a P-P plot represent the values corresponds to different cumulative probabilities in the empirical distribution and theoretical distribution respectively. For both Q-Q plot and P-P plot, a straight line with a slope of one indicates the two datasets exactly follow the same distribution. CDF represents the cumulative distribution function. Similarly, a straight line with slope one in the third subplot 'Empirical and theoretical CDF' indicates that the empirical cumulative distribution is the same as the theoretical one (beta distribution here). The Kolmogorov-Smirnov goodness-of-fit test shows we can't reject the null hypothesis that the predicted readmission risk follows a beta distribution [33].

We fitted beta regression to the predicted readmission risk using the same set of predictors used by the first stage model. For the discharge day readmission risk, the estimated risk from the first stage model instead of the actual outcome (0 or 1) was used, as beta regression doesn't model

extreme value 0 or 1 [33]. A pseudo-R-squared value of 0.88 indicates that the beta regression model fits the predicted readmission risk well.

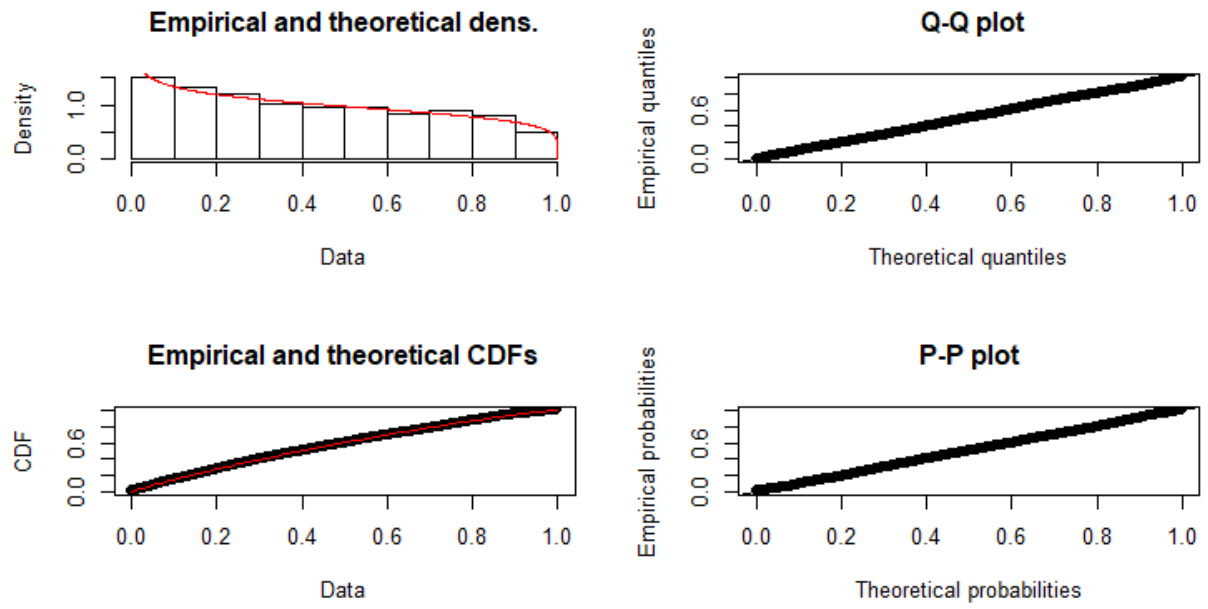


Figure 2.1 Results of fitting the predicted readmission risk using a beta distribution.

2.3.3 Unsupervised Clustering Results

After fitting a beta regression model on the daily readmission probabilities estimated by the first stage model using patients-day data as predictors, we used the fitted beta regression model for predicting the daily readmission risk. Then, we performed K-mean clustering on the daily readmission risk predicted by the beta regression model. Patient groups based on 30-day

readmission risk trends over time may be seen in Figure 2.2. Four patient groups with different readmission risk trends were identified. The trends over 5-days (at 6 a.m.) with the first probability generated at admission and the last at discharge is depicted. We identified four patient groups based on unsupervised cluster analyses. The 'decreasing risk' cluster had 131 (24.5%) encounters. Its average readmission probability decreased from 0.69 at admission to 0.30 at discharge. These patients entered the hospital with a relatively high 30-day readmission risk that decreased substantially (3 times) by hospital discharge. The remaining patient groups maintained a more consistent readmission risk. This included the 'high risk' group of 113 (21.2%) encounters with average readmission probabilities maintained above 0.75 over their course of care. Alternatively, the 'low risk' cluster with 113 (21.2%) encounters was admitted with a relatively low 0.39 probability of readmission that decreased to 0.21 at discharge. The 'Moderate' cluster had 177 (33.1%) encounters and its average readmission probability remained around 0.61.

We conducted the non-parametric Kruskal-Wallis test to determine whether a predictor is discriminative between the four clustering groups. The results show that 18 predictors do not have equal means across the four clusters ($p < 0.0001$). The distribution of these discriminative predictors is shown in Table 3.

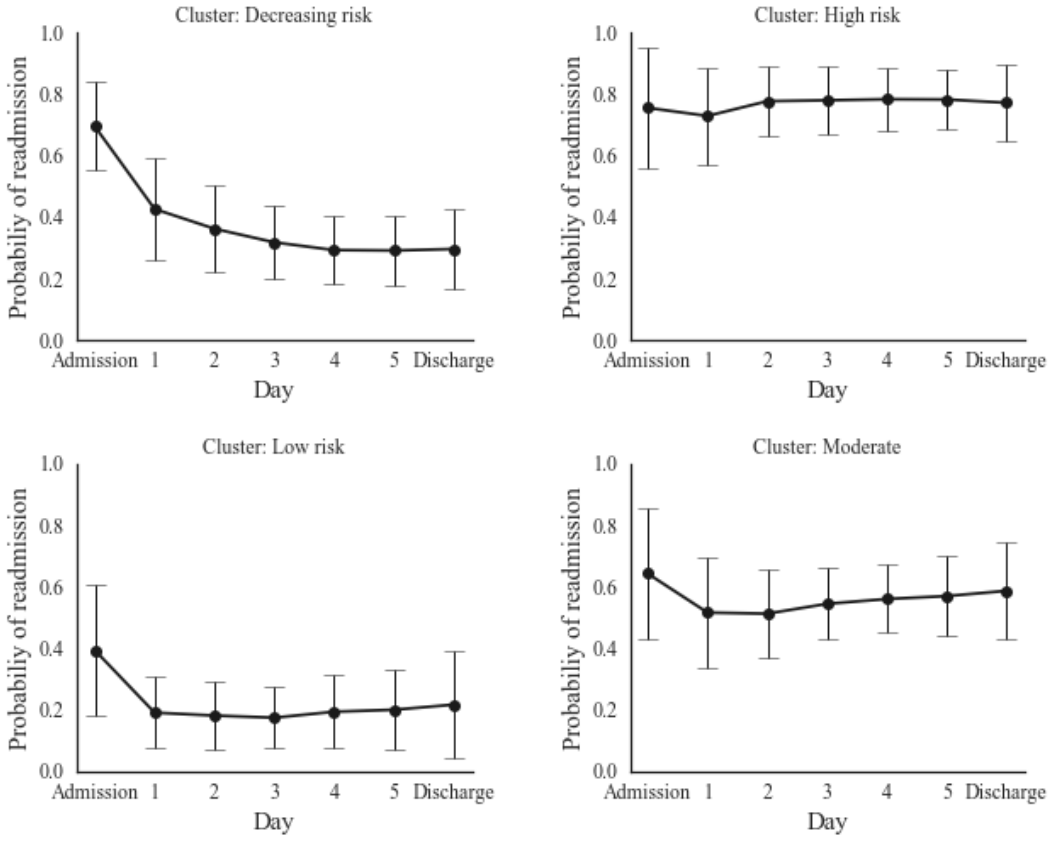
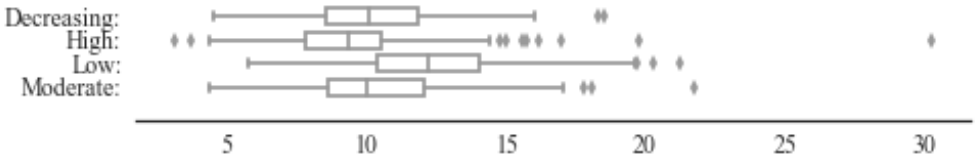
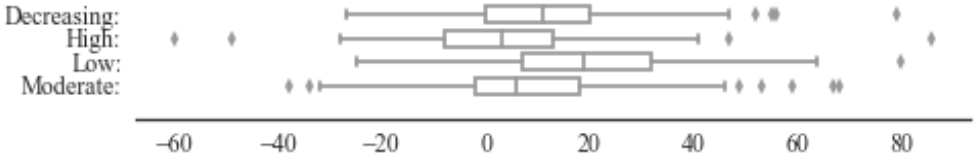
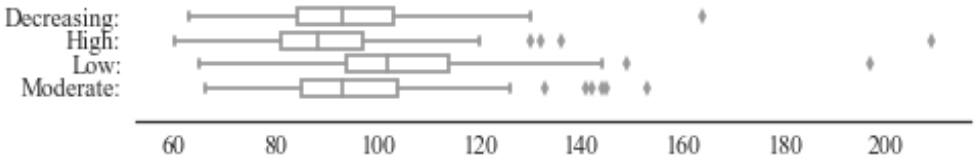
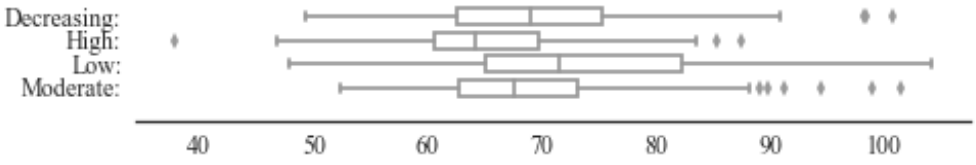
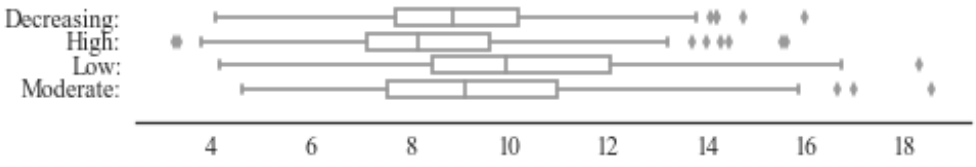
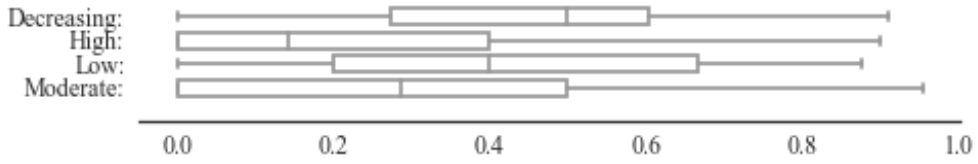
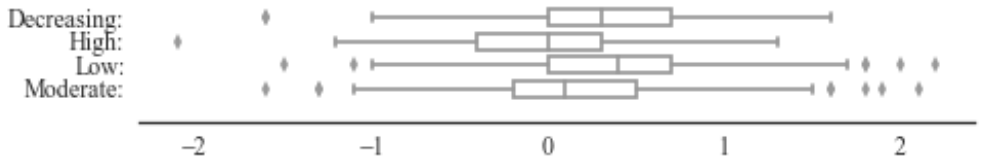
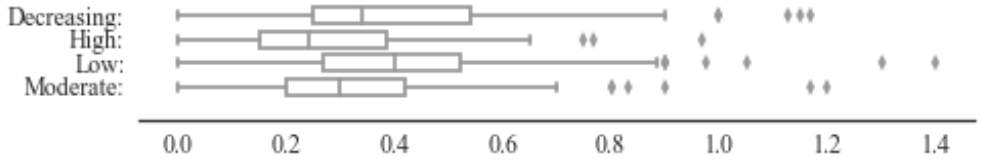
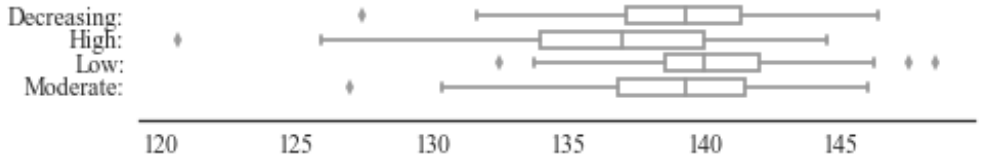
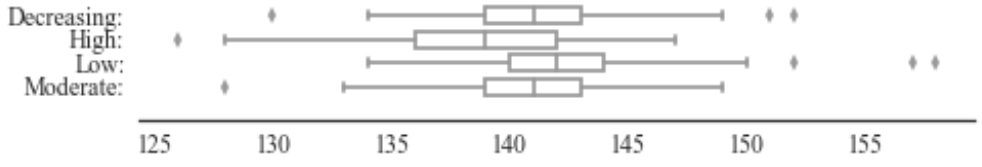
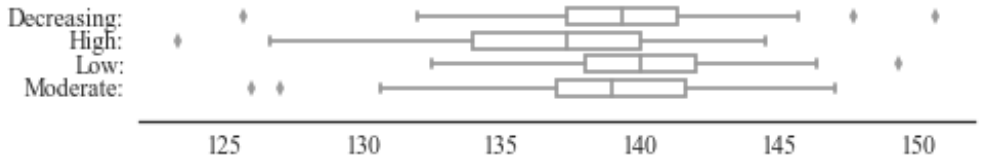
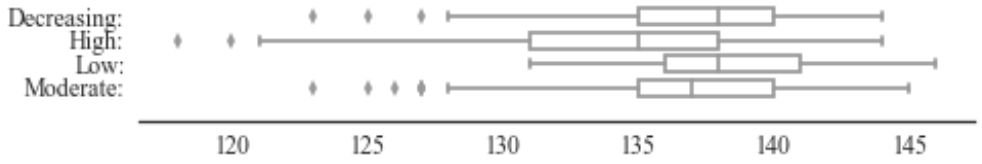
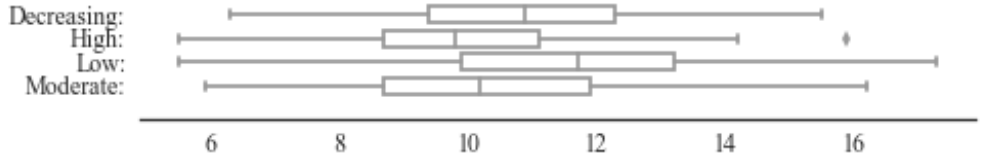
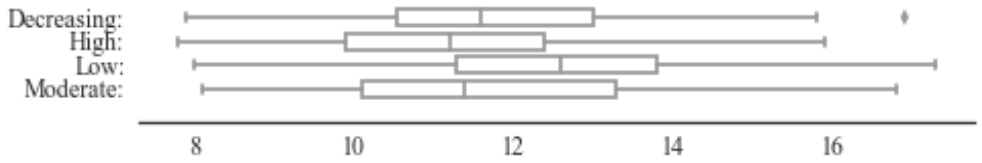
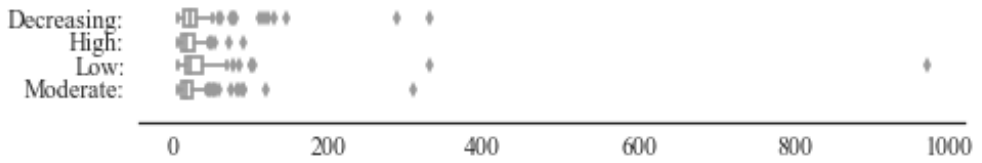


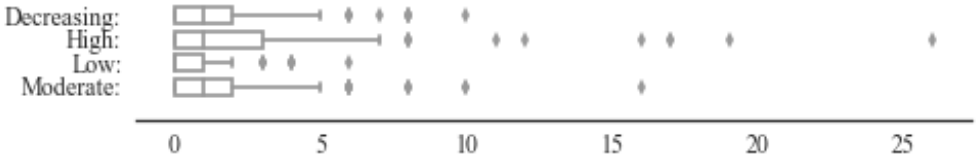
Figure 2.2 Daily readmission probabilities for four clusters of patient encounters. 30-day readmission probabilities are shown from admission to day 5 in addition to the patients' discharge day. Admission represents the time from when the patients arrived at the emergency department to the time when they were admitted. In each plot, the thick black line represents the mean 30-day readmission probability for all encounters in that cluster. The error bar represents one standard deviation from the mean value.

Table 2.3 Summary of discriminative predictors for each cluster, shown as boxplots. These predictors were produced by the Kruskal-Wallis test at a significance level of 0.0001.

Predictor	Boxplot
Diastolic blood pressure:	
Standard deviation (mmHg):	
First minus last value (mmHg):	
Maximal value (mmHg):	
Average value (mmHg)	
Average absolute change (mmHg)	
Potassium:	

Normalized index of minimal value:	
First minus last value (mmol/L)	
Average absolute change (mmol/L)	
Sodium:	
Average value (mmol/L)	
Maximal value (mmol/L)	
Average of last 3 values (mmol/L)	

Minimal value (mmol/L)	
Hemoglobin:	
Minimal value (gm/dL)	
Average value (gm/dL)	
Average of last 3 values (gm/dL)	
Maximal value (gm/dL)	
ALT:	
Minimal value (Units/L)	
Number of prior hospital visits:	



From Table 3, we can see that most of the discriminative predictors are laboratory test predictors of potassium, hemoglobin, sodium, and ALT. The other discriminative predictors are diastolic blood pressure and the number of prior hospital visits.

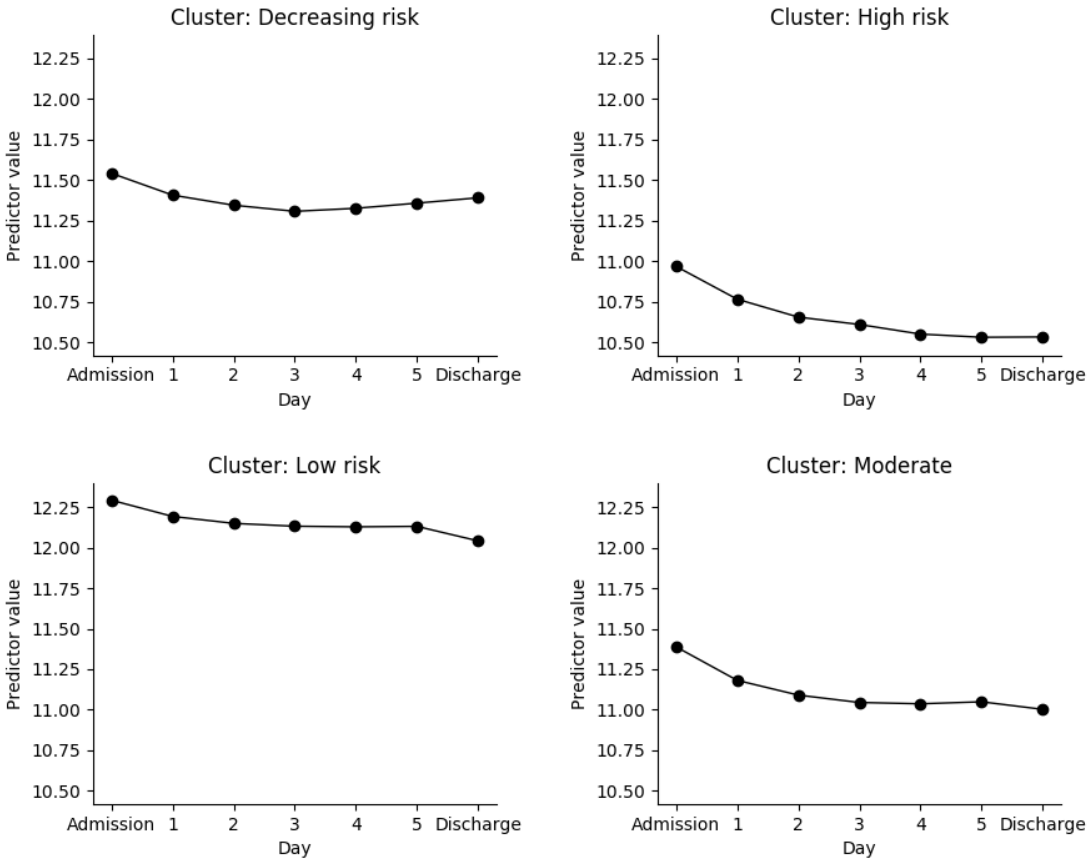
Overall, a lower hemoglobin level is associated with higher readmission probability during the stay, consistent with the fact that a lower hemoglobin level indicates anemia and increased risk of readmission. Larger decreases in potassium and diastolic blood pressure levels are associated with a decrease of readmission probability. The minimal potassium level occurring later during hospitalization is associated with decreasing readmission probability. Also, lower level of diastolic blood pressure and sodium are associated with higher readmission probability.

A larger decrease in diastolic blood pressure is associated with a lower readmission probability. This probably reflects the condition of patients with diastolic heart failure who have an increased diastolic blood pressure [36]. Higher variability, higher maximal and average value of diastolic blood pressure are associated with encounters whose readmission probabilities stay low, while lower values for these predictors are associated with encounters whose readmission probabilities remain high during hospitalization.

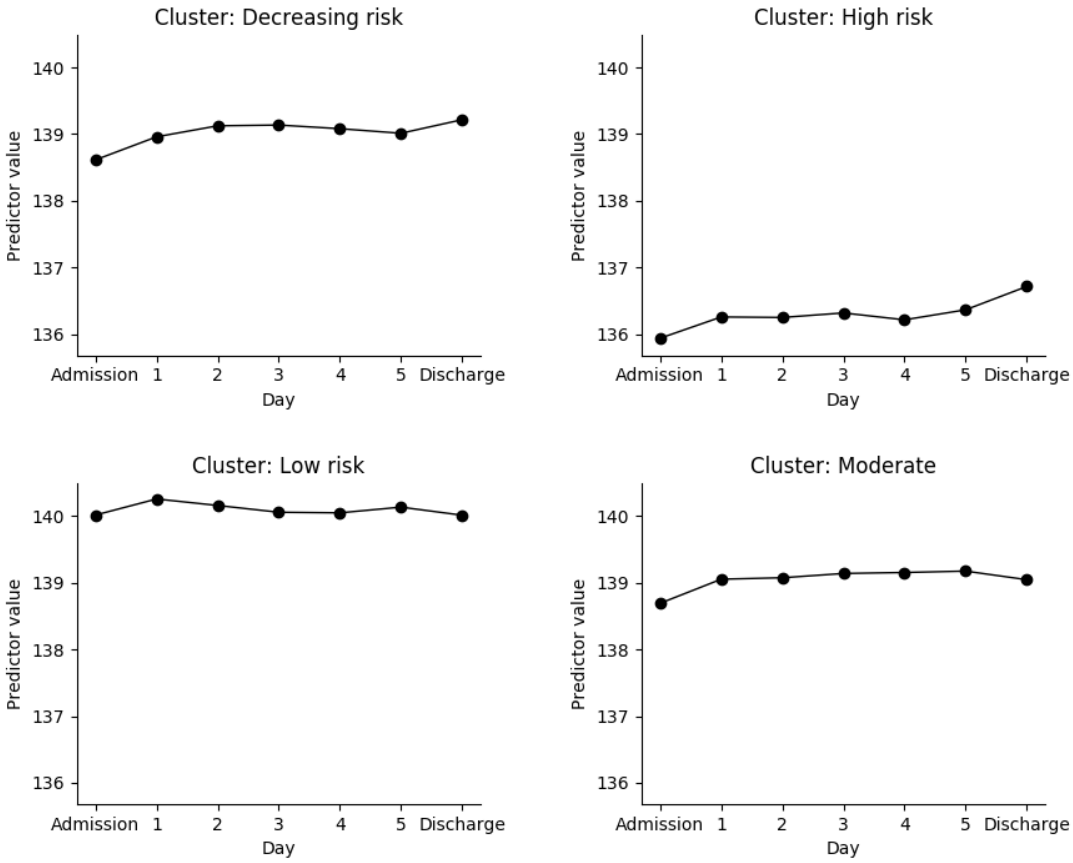
An important discriminative administrative predictor is the number of prior hospital visits. Specifically, a higher number of prior hospital visits is associated with a higher readmission probability.

To further investigate how the above discriminative predictors are associated with readmission risk patterns, we showed how those predictors change with time within each of the four readmission risk groups. As there are many predictors, given limited space, we only showed results of the main features we discussed above (hemoglobin, sodium, potassium, diastolic heart failure) in Figure 2.3. The results for the complete list of the 17 dynamic discriminative predictors are shown in Figure 7.1 in the supplemental material.

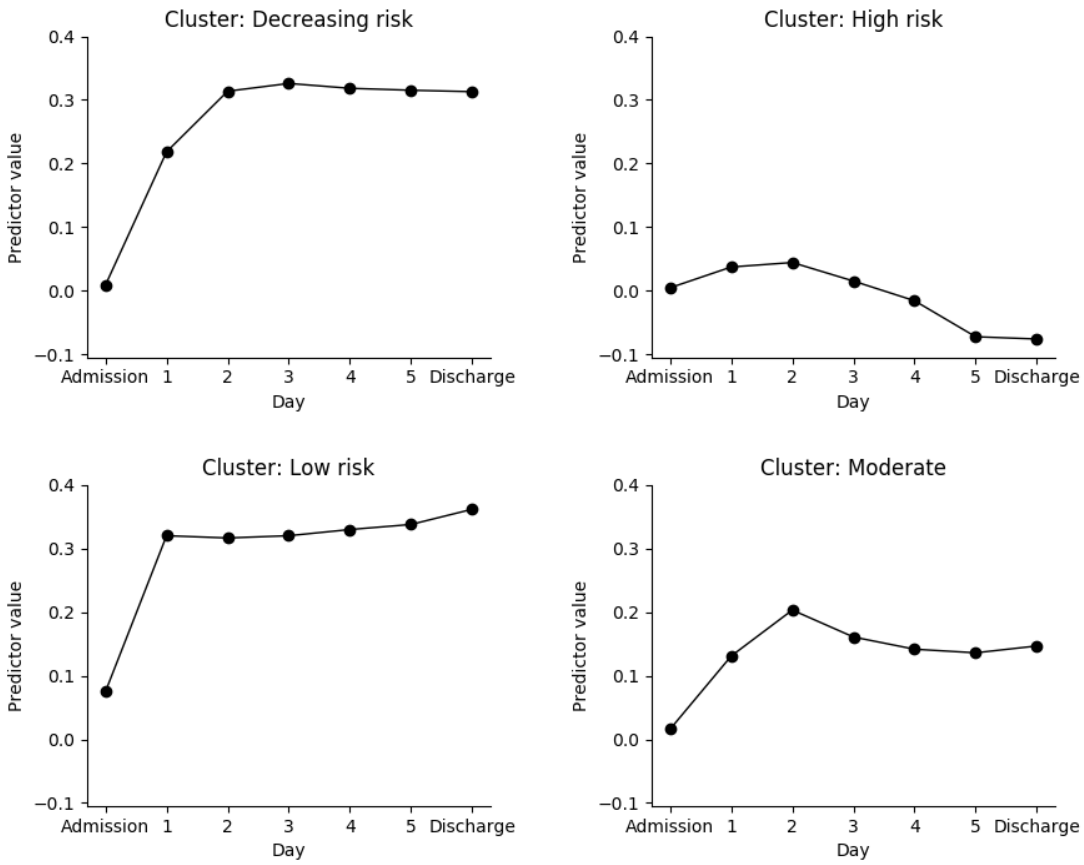
From Figure 2.3, it's clear to see that: a lower and decreasing level of HGB is associated with higher readmission risk; a lower level of sodium is associated with higher readmission risk; a larger decrease of potassium level from admission and minimal potassium level occurring closer to discharge are associated with lower readmission risk; a larger decrease of diastolic blood pressure from admission is associated with lower readmission risk.



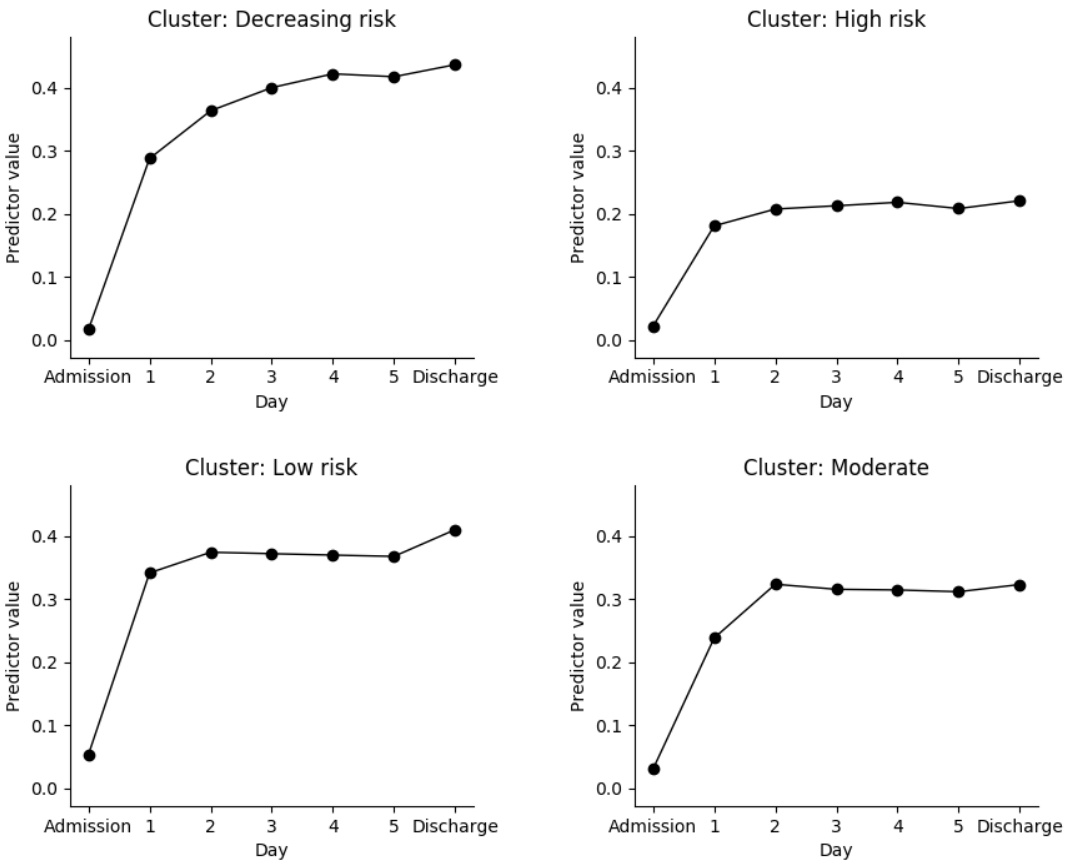
(a) Average value of last three hemoglobin (gm/dL)



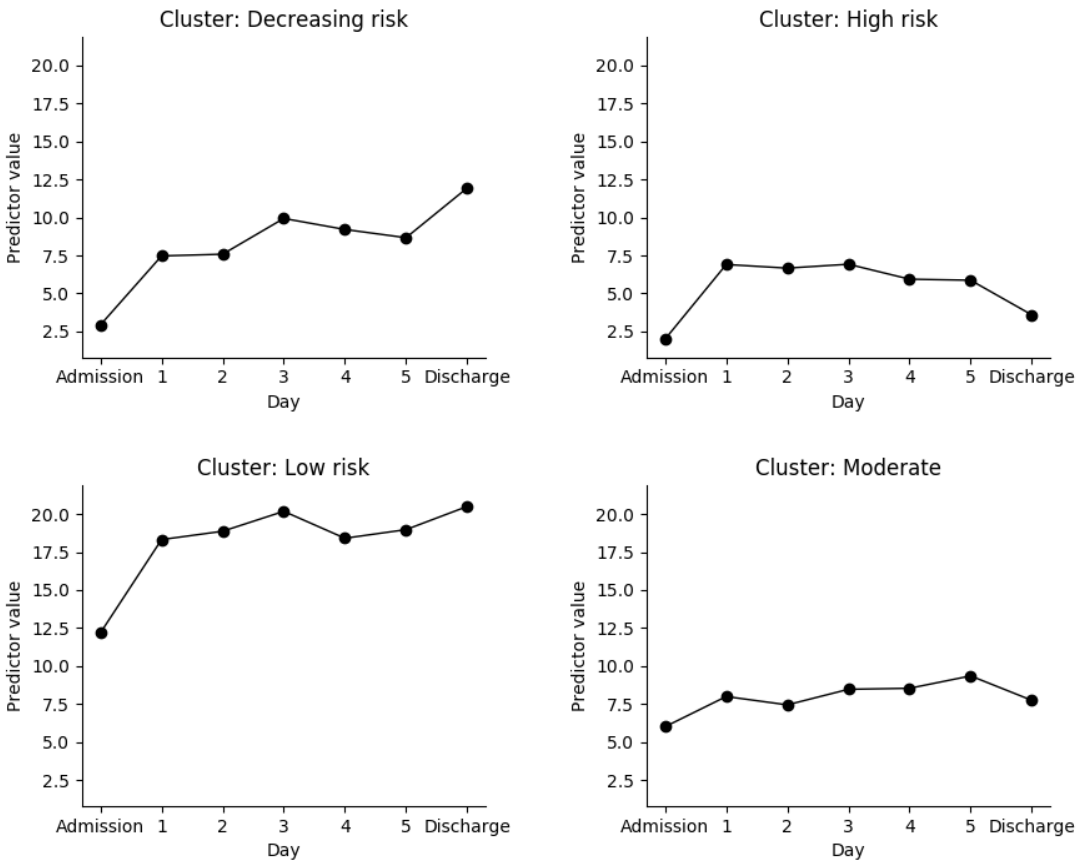
(b) Average value of last three sodium (mmol/L)



(c) Decrease of potassium level from admission (mmol/L)



(d) Normalized time of minimal potassium starting from admission

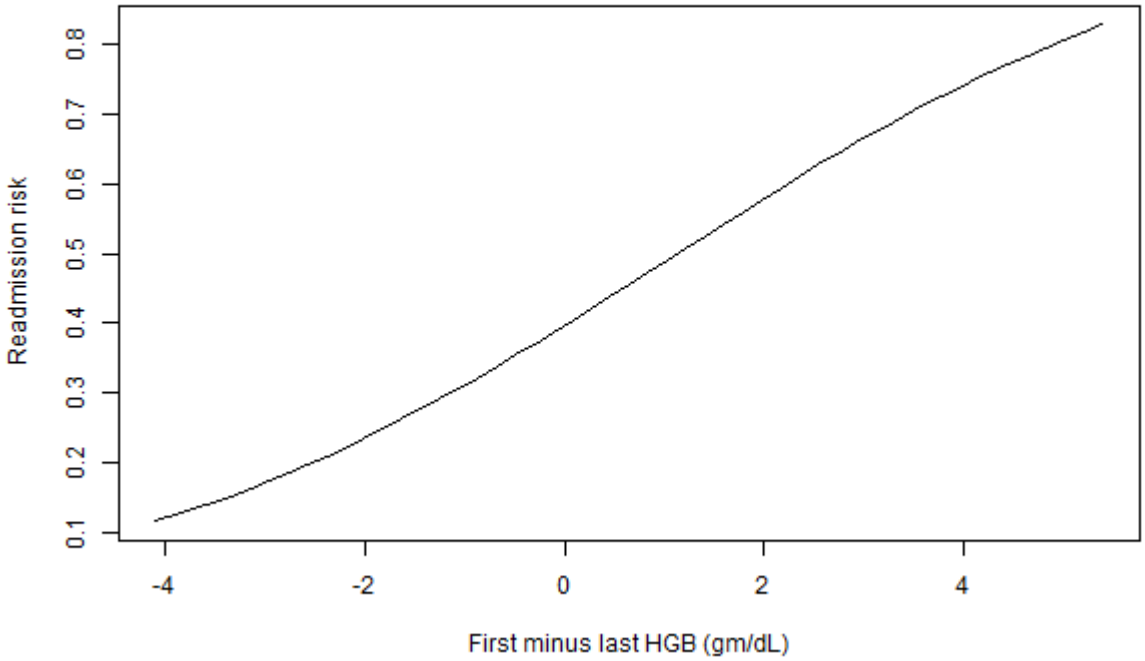


(e) Decrease of diastolic blood pressure level from admission (mmHg)

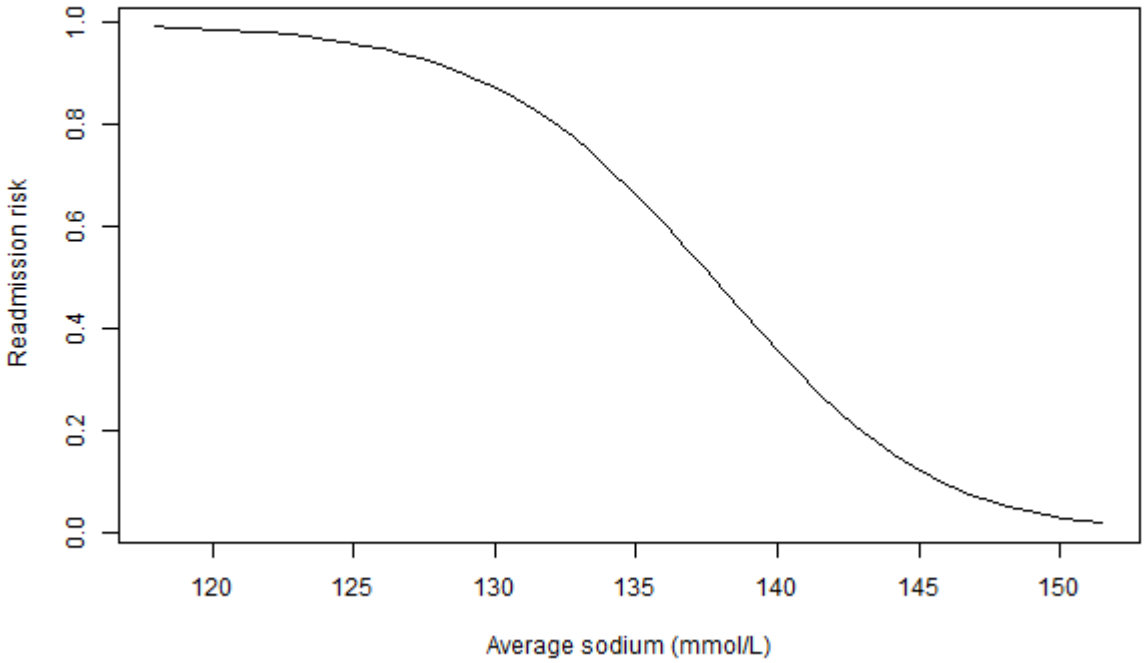
Figure 2.3 Change of discriminative predictors values over time from admission to discharge within each patient risk group.

2.3.4 Partial Dependence of Predictors

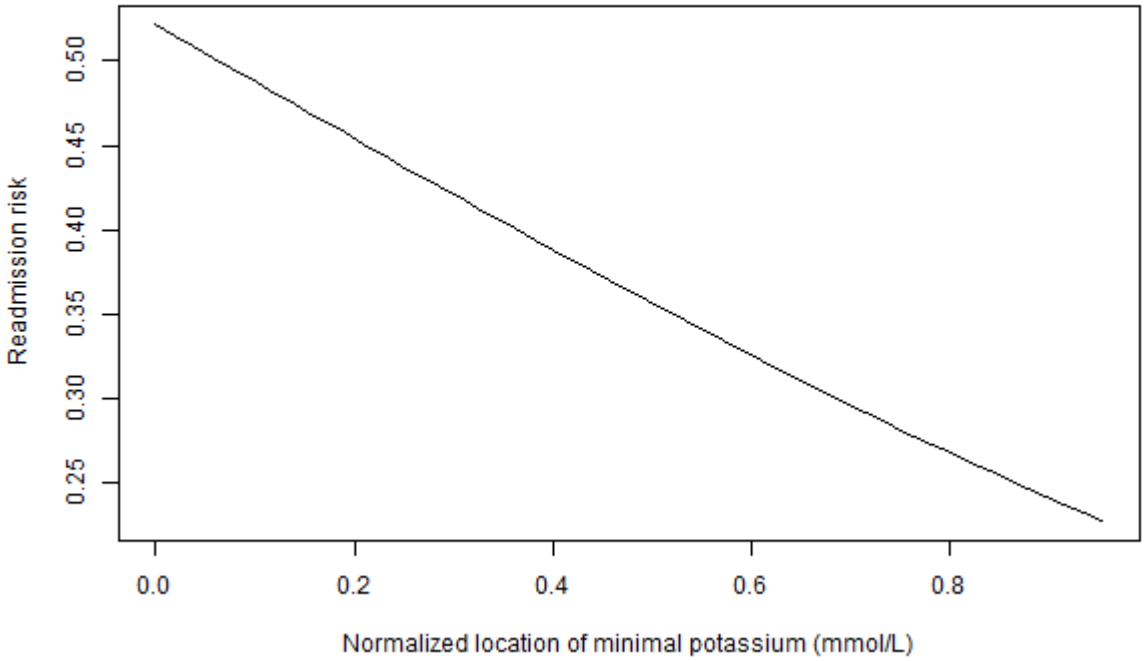
Next, we show the partial dependence plots for the main predictors discussed above in Figure 2.4: hemoglobin, sodium, potassium, diastolic heart failure. Those partial dependence plots match with the results we see in the above section. Precisely, a larger decrease of hemoglobin from admission is associated with higher readmission risk. As we previously discussed, a low hemoglobin level indicates bad patient condition regarding blood oxygen level and was found to indicate higher readmission risk. A larger decrease of hemoglobin means a lower level of hemoglobin at discharge, leading to higher readmission risk. A higher average value of sodium over patients' stay is associated with lower readmission risk. Unlike hemoglobin, the marginal effect of sodium on readmission risk decreases when the level of sodium is on the low or high end of its range of values in our patient cohort (120-150 mmol/L). Again, the minimal potassium occurring at a later time of the patient's stay is associated with a lower readmission risk. A larger decrease of diastolic blood pressure level is associated with lower readmission risk. These findings seem to indicate that a more stable patient condition at discharge is associated with a lower readmission risk, which is consistent with intuition.



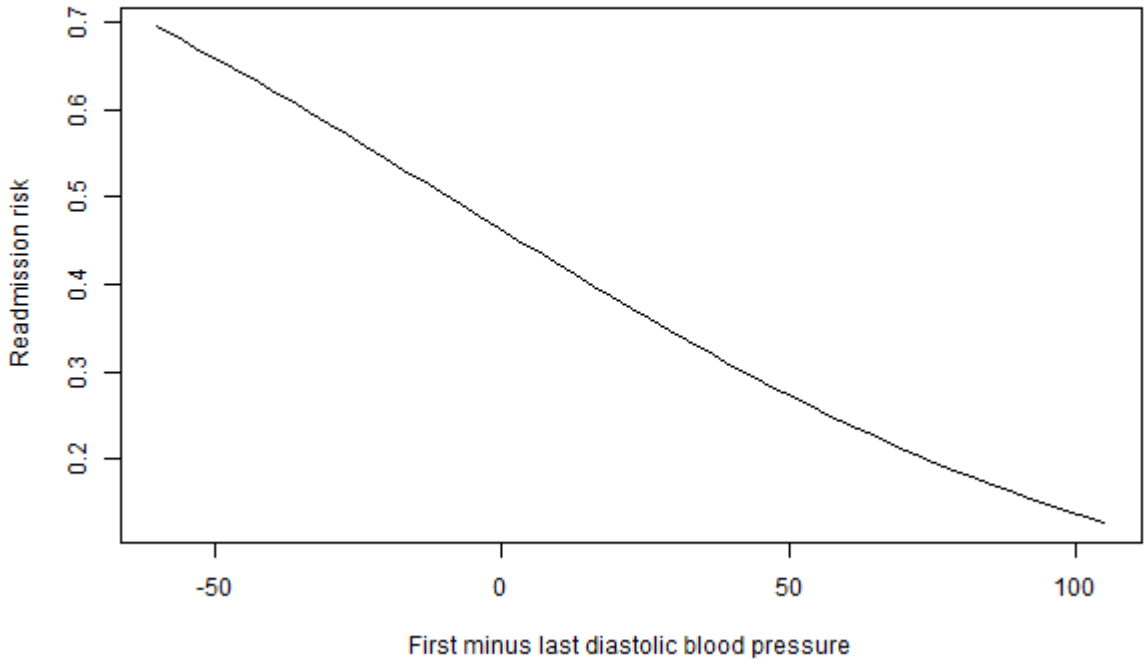
(a) Decrease of hemoglobin from admission (gm/dL)



(b) Average value of sodium (mmol/L)



(c) Normalized time of minimal potassium starting from admission



(d) Decrease of diastolic blood pressure (mmHg)

Figure 2.4 Partial dependence plots for the main predictors: hemoglobin, sodium, potassium, diastolic blood pressure.

2.4 Discussion

We created a two-stage daily readmission risk prediction model for a cohort of patients with a primary diagnosis of HF. Further, we used clustering analysis to analyze how the likelihood of readmission changes over time, and learn about the discriminative features that explain different

trends. There are two major contributions to this work. First, we built a daily readmission risk prediction model that can be used in practice using a two-stage modeling approach. At the first stage, we applied logistic regression model to estimate the counterfactual daily readmission risk. At the second stage, we utilized a beta regression model to predict the counterfactual daily readmission risk using patients-day data. The fitted beta regression model is used as the final risk prediction model. Second, we found predictors that indicate different dynamic readmission risk patterns during hospitalization.

A larger decrease of hemoglobin and diastolic blood pressure is associated with a higher readmission probability. Many studies have shown that reduced hemoglobin in congestive heart failure patients is associated with increased risk of hospitalization and all-cause mortality [20][37][38][39][40][39][41][42][43]. Our study result is concordant with those studies.

For BUN, the effect on readmission risk may not be the magnitude of the change, but rather the decrease from a high BUN measurement at admission. A higher level of the BUN is associated with higher risk of all-cause mortality for decompensated heart failure patients [44]. In addition, decreased renal function during hospitalization is positively associated with increased HF hospitalizations [45][46]. Those with a larger decrease likely started high. A larger decrease of BUN probably reflects the improvement of patient outcome and, as a result, lower readmission probability.

An odds ratio of 0.1 for the normalized index of minimal potassium means that the readmission probability will decrease by 90% if the minimal potassium occurs as the last measurement instead of the first measurement, given the other predictors remain fixed. Similarly, minimal sodium occurring later during hospitalization indicates a lower readmission probability. Moreover, patients' average sodium levels are negatively correlated with readmission probability.

By examining the changes of readmission probabilities over the course of a patient's hospital stay, we discovered distinct readmission risk groups. The clusters provide additional insights about the discriminative predictors, which are difficult to detect solely by developing a static prediction model. The findings from the clustering analysis are supported by evidence either in the literature and our static prediction model. For example, both the literature and the prediction model show that number of prior hospitalization is positively correlated with risk of readmission [17][47][48][49][50].

Several findings regarding the laboratory test predictors have not been reported in the literature. These findings provide evidence that temporal patterns of these measurements are relevant in predicting the risk of HF patient readmissions. In particular, a higher level of hemoglobin, a larger decrease in value of potassium and diastolic blood pressure from admission to discharge indicates a lower readmission risk. A higher average value of last three hemoglobin measurements, which can be considered as patients' stability measures close to discharge indicates lower readmission risk [47]. This finding agrees with Nguyen et al.'s study for general readmission

[26]. They found that vital sign instability on discharge is associated with increased risk-adjusted 30-day mortality and readmission rates.

Given these findings, further prospective analyses can be designed around developing guidelines such as the following to prevent discharging patients with high readmission risk:

1. Provide special attention to patients who have prior hospital visits.
2. Measure the patients' diastolic blood pressure level immediately after admission and immediately before planning to discharge them. Our results suggest that patients have a high readmission risk if the last diastolic blood pressure level is abnormal or the decrease in patients' diastolic blood pressure from admission to discharge is less than or equal to 4.
3. Measure the patients' hemoglobin level immediately before planning to discharge them. Check the value of the last hemoglobin measurement and the average value of last three hemoglobin measurements. Our results suggest the patients have a high readmission risk if the values are lower than normal level.
4. Measure the patients' potassium level immediately before planning to discharge them. Our results suggest that patients have a high readmission risk if the last potassium measurement is above the normal level.

5. Monitor the trajectory of the following dynamic discriminative predictors: average value of last three HGB and sodium, a decrease of potassium and diastolic blood pressure level from admission, and normalized time of minimal potassium starting from admission. Mark patients whose last three HGB and sodium are low or decreasing and whose potassium and diastolic blood pressure level remains high and doesn't decrease as potential high readmission risk patients.

Even though our prediction model has a relatively high predictive power and is the first to have identified the different dynamic readmission risk patterns, there are several limitations, and our findings need to be interpreted in the context of these limitations.

First, we conducted the clustering analysis of patients' readmission risk by using readmission probabilities under a normalized period of stay. The normalization had its advantages in that we could compare readmission probability profiles across patients, but the disadvantage is we lose out on details that could help with the prediction. Ideally, we want to perform clustering on readmission risk by controlling for the entire duration of their stay. However, this is impossible due to different lengths of stay of different patients. One option could have been to use warping methods such as dynamic time warping (DTW), which matches two time series and yields a minimal distance assuming the two time series only differ in speed. But applying DTW would have meant that we could not have interpreted the results in a medical context. Nonetheless, we are still able to see the four different dynamic readmission risk patterns from the clustering

analysis. We believe clustering readmission risk over longer patients' stays won't qualitatively change our findings as patients' readmission risk tends to vary less towards later of their stay.

Second, we treated each encounter instead of each patient as a sample in our cohort for the first stage model. The assumption that our sample is independent and identically distributed (i.i.d.) is likely violated because multiple encounters may correspond to one single patient and those encounters should be non-independent. For predictive modeling, the i.i.d. assumption is important concerning model prediction performance evaluation. As we randomly split data into training and test data, we require the sample to be i.i.d. to avoid biased performance evaluation. The reason we used encounters instead of patients is twofold: 1) we would lose 10.5% of our current sample if we treat each patient as a sample (478 patients and 534 encounters), and 2) we believe it's okay to assume that two encounters from the same patient are independent given that most of our predictors are clinical data. We hope an analysis on using data from repeated visits from the same patient for predictive modeling becomes an opportunity for future research. Another future study could be a different modeling approach such as a longitudinal model that captures the correlation between the repeated visits of the same patient.

Third, we were not able to exclude planned readmissions from our training data because we couldn't identify them. We were mostly interested in predicting unplanned and avoidable readmission. Having planned readmission in our datasets may have introduced bias into the

training data and affected the training process. To make our prediction tool more useful, we can exclude planned readmissions in the future analysis when the data become available.

Finally, we conducted our analysis of dynamic readmission risk patterns retrospectively. A retrospective study is more prone to bias and confounding issues compared to a prospective study. A prospective approach should be taken to study the readmission risk patterns over patients' stay in the future.

Given these limitations, the difficulty of accurately predicting heart failure readmission in general, and the complex nature of the problem itself, we believe studying readmission risk requires a more holistic approach than simply analyzing EHR data. We should study many aspects of healthcare for heart failure patients, from patient treatment, post-treatment care, to readmission intervention strategies as an integrated system.

2.5 Conclusion

We applied supervised machine learning algorithms in a two-stage approach and built a daily readmission risk prediction model that can be used in practice to predict the risk of 30-day readmissions for HF patients. To deal with unknown counterfactual daily readmission risk, we applied a logistic regression model trained on encounter data to estimate the counterfactual daily risk using patients-day data. A beta regression model is further fitted to the counterfactual daily risk using patients-day data, which is used as the final daily readmission risk prediction model.

We conducted clustering analysis and found groups of encounters with different readmission risks. We also identified the discriminative features associated with risk of readmissions. Notably, we found a lower decrease in value of Potassium and diastolic blood pressure, a lower value of hemoglobin close to discharge is indicative of high risk of readmissions. These can be used as instability indicators to identify patients with high risk of readmissions. As a result, intervention can be provided to these patients before they are discharged. For future study, besides collecting more data, we propose that controlled experiments or causal inference can be conducted to study the discriminative features and the reasons for readmissions. To improve the performance of the prediction model, we also propose to exclude planned readmissions from the future study if given more data available.

Chapter 3

Xerostomia Prediction and Knowledge Discovery

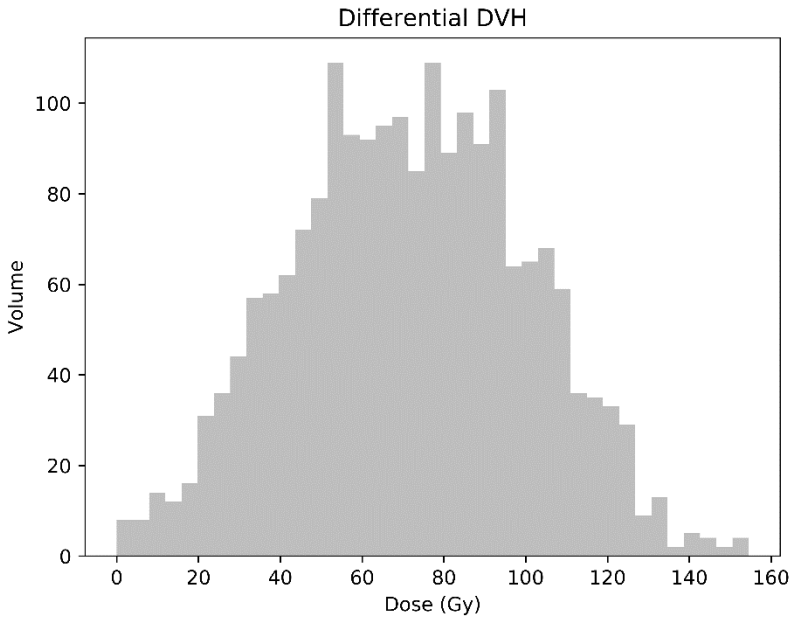
3.1 Introduction

Head and neck cancer (HNC) patients are commonly treated with radiation therapy (RT). However, the anatomic complexity of this part of the body increases the risk of normal tissue injury. Of the spectrum of side-effects due to the injury to organs at risk (OAR) from RT, xerostomia arising from parotid and submandibular glands radiation have received significant study. With modern intensity modulated radiotherapy, it's now well established that the mean PG and SMG radiation dose is associated with the risk of developing xerostomia [51–54], providing an opportunity to modify these radiotherapy techniques in the management of HNC patients.

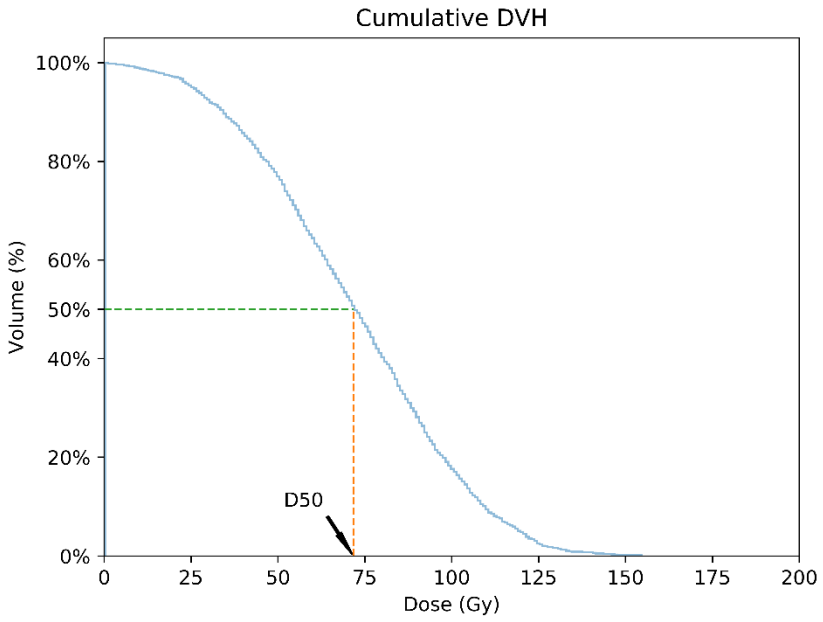
Despite these technological advancements, RT-induced xerostomia continues to be a significant clinical challenge commonly affecting the patient-reported quality of life. To study RT-

induced xerostomia, most existing literature used aggregated or summarized dose features within certain organs, such as mean dose, and dose-volume histogram (DVH) features in PG [51–53,55].

DVH is a histogram that describes the relationship between radiation dose levels and the corresponding tissue volumes of a radiation treatment plan. It has been commonly used to describe three-dimensional radiation dose in a two-dimensional figure and compare different radiation treatment plans. There are two types of DVH figures: differential DVH and cumulative DVH. Typically, a cumulative DVH is used to describe the dose distributions. Differential DVH is a normal histogram of the dose levels in voxels (volume element), where the y-axis is the number of voxels that corresponds to a certain dose level. Cumulative DVH shows the volume in the percentage of the entire volume of the target region whose dose level is higher than a certain value. From the cumulative DVH, we can derive different dose features such D50, D90, which is the dose level that 50% of the volume received no less than. For D90, 90% of the volume received a dose no less than D90. Figure 3.1 shows an example of the differential DVH and cumulative DVH.



(a) A differential dose-volume histogram



(b) A cumulative dose-volume histogram

Figure 3.1 Example of a different dose-volume histograms and a cumulative dose-volume histogram.

The disadvantage of using summarized dose feature within certain organs is that it loses the spatial information for the radiation dose within an organ: different spatial distributions of dose within an organ can yield the same mean dose and DVH features.

On the other hand, the spatial information of dose is the key to understanding the local dose effect on xerostomia. Pre-clinical investigations suggest that RT-induced xerostomia may not only be related to the PG dosimetry but that the spatial location of the subvolume of the PG that is secondarily irradiated is particularly important [56,57]. In rat models, irradiation of the caudal

portion of the PG caused not only xerostomia but was associated with salivary function recovery in contrast to irradiation of the cranial portion. These investigators have subsequently demonstrated that this recovery may be related to the presence of stem/progenitor cells that are responsible for the recovery of radiation-induced xerostomia [58].

Data mining investigations by our group within our Oncospace informatic infrastructure have demonstrated human evidence of the low dose impact of PG radiation [59–61] causing xerostomia at 3-6 months. As our head and neck conformal radiotherapy typically utilizes coplanar beam arrangements, we hypothesized that this might represent the low dose irradiation of the cranial portion of the PG that occurs when the target volume extends more superiorly. Pilot investigations by our group suggested that the cranial half of the PG and its dosimetry may be more important in causing severe xerostomia in our HNC patients [61].

To more robustly evaluate the influence of specific subvolumes of the PG and SMG on injury and symptoms of severe xerostomia after RT, we applied supervised machine learning methods and a radio-morphology approach using voxel-based dose features to predict for parotid-injury-causing acute xerostomia. Radio-morphology parametrically represents the spatial dose distribution within normalized anatomic structures using voxel-based or shaped-based dose features, which are consistent across patient cohort [62]. The supervised machine learning method ultimately used can learn the influence of spatial dose pattern across organ sub-volumes on xerostomia by assigning higher weights to the voxel dose features of the predictive regions and

lower or zero weights to the less predictive regions. We define the spatial pattern of the learned weights distributed across voxels as the voxel importance pattern.

3.2 Methods and Materials

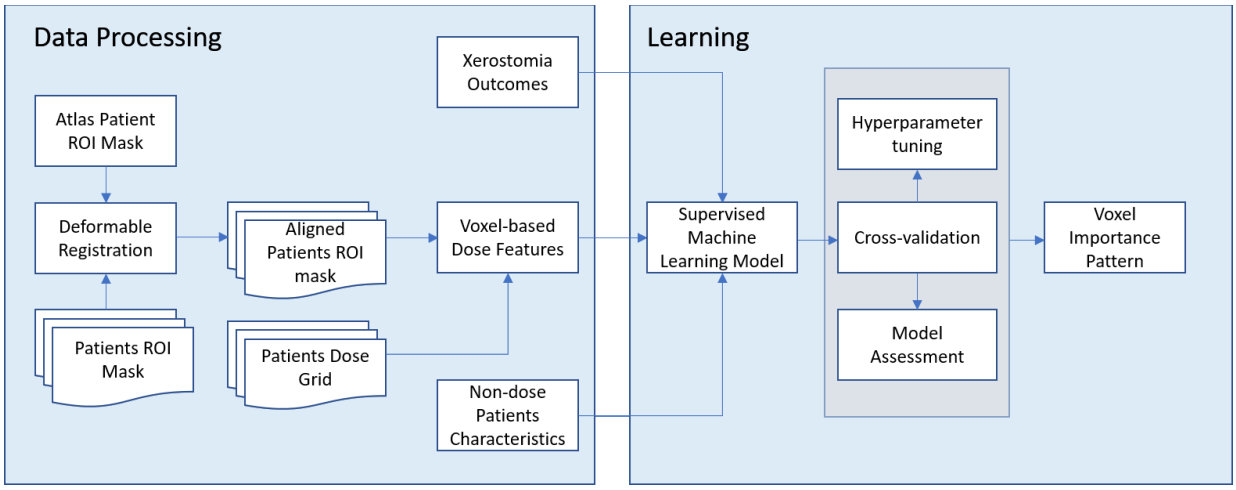


Figure 3.2 The flowchart of the key steps for this analysis. (ROI: region of interest)

3.2.1 Patients

Our study population included 427 HNC patients who were treated with parotid-sparing intensity-modulated RT with curative intent from January 2008 to December 2016 and for whom xerostomia scores were available in our database. The data included both single-sided and double-sided neck treatments. All patients were seen weekly during radiotherapy for on-treatment visit assessments and in follow-up typically every 3-4 months for the first three years and every six months after

that. All study assessments, including the National Cancer Institute Common Terminology Criteria for Adverse Events (CTCAE) xerostomia grading, were performed prospectively at the point of care during routine treatment and follow-up visits. The study cohort excludes patients who do not have all PG and SMG contoured to ensure complete radiation dose features in these two major salivary glands.

3.2.2 Features

The features are categorized into two parts: planned radiation dose features, patients' demographic and clinical pathology features. To capture the spatial information of radiation dose explicitly, organs are sampled into voxels, and actual dose in the voxels are the radiation dose features. The voxels are defined as uniformly distributed points in organs downsampled from the original dose grid. Specifically, we used voxel-based dose in PG and SMG.

Patients' anatomical structures are spatially different. To obtain consistently identifiable dose features for all the patients, we aligned patients' structures to a common reference frame using a deformable registration technique, i.e., the Coherent Point Drift algorithm [63]. Further, to consider the factor of tumor location in the dose features, we mirrored the patients' structures so that the spatial relationship of the organs, which we derived the dose features from, is not left versus right, but rather ipsilateral versus contralateral relative to the disease side. The framework and actual steps to generate the dose features were described in a prior study [62].

The socio-demographic and clinical pathology features included are gender, race, age, attending physician, baseline xerostomia grade, tumor characteristics (TNM stage), chemotherapy: yes/no, human papillomavirus (HPV) status: positive/negative, whether feeding tube was used, and tumor site. Tumor site was characterized by mapping ICD-9 and ICD-10 diagnosis codes to specific site definitions. These features were chosen to capture the basic patients' characteristics and factors that may relate to xerostomia. There is no missing data in continuous features, and missing data in categorical features were treated as a new "missing" category. Therefore, no missing data imputation was performed for features in our study.

3.2.3 Outcome Measure

The primary outcome measure was the CTCAE xerostomia grading at the start of the first follow-up period, which is three months after RT. We choose to look at xerostomia three months after RT as an indication of acute xerostomia caused by radiation-induced injury (as opposed to the inclusion of recovery). For supervised predictive modeling, a binary classification problem was created by grouping the xerostomia grading into two categories: severe xerostomia if the grade is 2 or 3 and no/mild xerostomia if the grade is 0 or 1. The prediction target is whether a patient will develop severe xerostomia three months after RT. Baseline xerostomia grade was measured either before treatment starts or during the first week of treatment. For the patients who dropped out before three months post-treatment, last xerostomia measure was carried forward. Patient cohort characteristics at baseline stratified by xerostomia prediction outcome are shown in Table 3.1.

Table 3.1 Patient characteristics (N= 427) at baseline. Summary statistics for continuous variables (indicated by '*') include the mean and interquartile range. Summary statistics for categorical variables is the count and percentage value. *p*-value is obtained for the two-sample test.

Predictor	**Xerostomia grade ≥ 2 at 3 months post-RT**		***p*-value**
	No (N=282)	Yes (N=145)	
Age*	58.62 (52, 67)	58.47 (53, 64)	0.77
Gender			0.61
Male	210 (74.47%)	112 (77.24%)	
Female	72 (25.53%)	33 (22.76%)	
Race			0.24
Caucasian	196 (69.50%)	113 (77.93%)	
African American	65 (23.05%)	22 (15.17%)	
Asian/Pacific islander	8 (2.83%)	7 (4.83%)	
Other	13 (4.61%)	3 (2.07%)	
Attending physician			0.22
1	143 (50.71%)	69 (47.59%)	
2	58 (20.57%)	28 (19.31%)	
3	35 (12.41%)	25 (17.24%)	
4	3 (1.06%)	3 (2.06%)	
Missing	43 (15.25%)	20 (7.09%)	
Chemotherapy			< 0.01

Yes	198 (70.21%)	123 (84.83%)	
No	84 (29.79%)	22 (15.17%)	
HPV			< 0.01
Positive	185 (65.60%)	74 (51.03%)	
Negative	94 (33.33%)	71 (48.97%)	
Missing	3 (1.06%)	0 (0%)	
Feeding tube used			0.06
Yes	196 (69.50%)	88 (60.69%)	
No	83 (29.43%)	57 (39.31%)	
Missing	3 (1.06%)	0 (0%)	
Baseline xerostomia grade			< 0.01
0	216 (76.60%)	99 (68.28%)	
1	63 (22.34%)	33 (22.76%)	
2	3 (1.06%)	13 (8.97%)	
Primary tumor stage (T stage)			0.89
0	12 (4.26%)	6 (4.13%)	
1	50 (17.73%)	29 (20.00%)	
2	66 (23.40%)	44 (30.34%)	
3	47 (16.67%)	23 (15.86%)	
4	65 (23.05%)	35 (24.14%)	
Missing	42 (14.89%)	8 (2.84%)	

Regional lymph nodes stage (N stage)			0.11
0	69 (24.47%)	25 (17.24%)	
1	33 (11.70%)	25 (17.24%)	
2	128 (45.39%)	83 (57.24%)	
3	6 (2.13%)	5 (3.45%)	
Missing	46 (16.31%)	7 (2.48%)	
Distant metastasis stage (M stage)			0.51
Yes	16 (5.67%)	6 (4.14%)	
No	229 (81.21%)	132 (91.03%)	
Missing	37 (13.12%)	7 (2.48%)	
Tumor site			< 0.01
Oral cavity	52 (18.44%)	45 (31.03%)	
Oropharynx	54 (19.15%)	42 (28.97%)	
Nasopharynx	12 (4.26%)	14 (9.66%)	
Larynx	54 (19.15%)	17 (11.72%)	
Other	110 (39.01%)	27 (18.62%)	

3.2.4 Prediction Models

Three supervised machine learning algorithms were applied to our dataset: ridge logistic regression, lasso logistic regression, and random forest [31,64]. The 10-fold cross-validation area

under the curve (AUC) score for each model was compared after fitting each model to our dataset. The optimal hyperparameters for those supervised machine learning algorithms were chosen using 10-fold cross-validation while maximizing the AUC score on the hold-out data. Further, the cross-validation was repeated 40 times with different random splitting of the training data to ensure the learned hyperparameters don't depend on a specific random splitting of the training data.

To determine the best model, the voxel importance pattern obtained from each algorithm along with the prediction performance was evaluated among the three algorithms.

3.2.5 Voxel Importance Pattern

Lasso logistic regression learns a sparse set of features and assigns zero weight to non-important features. Ridge logistic regression doesn't learn a sparse solution but assigns larger weights to more important features. The magnitude of the feature weights learned by regularized logistic regression combined with hyperparameter tuning using cross-validation indicates the relative importance of the voxel-based dose features. Also, the radiation dose in different voxels is all on the same scale (average dose ranges from 9.09 Gy to 58.77 Gy), preventing the issue of having very different weights due to different feature scales. Therefore, the magnitude of the learned weights was used as a measure of voxel importance. The voxel importance indicates how much a change in the radiation dose in a voxel affects the probability of the patient developing xerostomia. Higher positive voxel importance indicates a larger chance of developing xerostomia if we increased the dose in that voxel.

The voxel importance for random forest was measured by computing how much the squared error decreased during the training process when partitioning data using a certain feature over all trees [31,32].

To visualize the voxel importance pattern, we normalized the learned weights from logistic regression and relative importance measure from random forest to be in the range of [0,1] and linearly mapped the normalized voxel importance onto the three-dimensional PG and SMG structure.

We further checked whether the voxel importance pattern is robust and consistent giving sample randomness. We randomly separated the data into five folds and repeated the analysis (the learning step in Fig. 1) five times on four folds. Then, the correlation coefficients of weights learned from the five analyses were computed. High correlation coefficients indicate that the voxel importance pattern is robust and consistent on different random samples.

3.3 Results

3.3.1 Acute Xerostomia

3.3.1.1 Acute Xerostomia Outcome

For the xerostomia prediction outcome, 145 (34%) patients have xerostomia (grade 2 and 3) three months post-RT. 89 (20.8%) of the patients dropped out before three months-post treatments in

the dataset. Last-observation-carried-forward was used to obtain the outcome for these patients. Figure 3.3 shows the distribution of xerostomia grade at baseline and three months post-RT. It also shows a large portion of patients developed xerostomia post-RT.

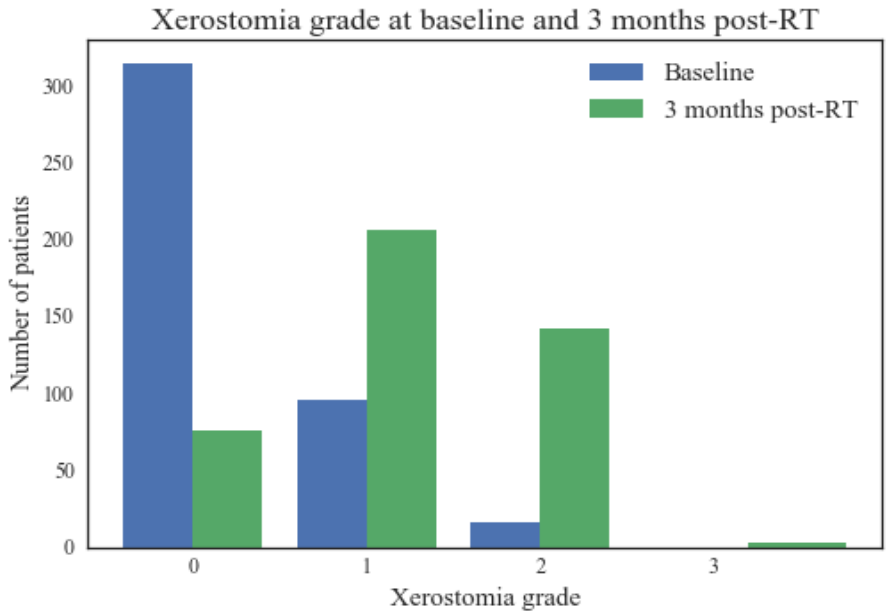


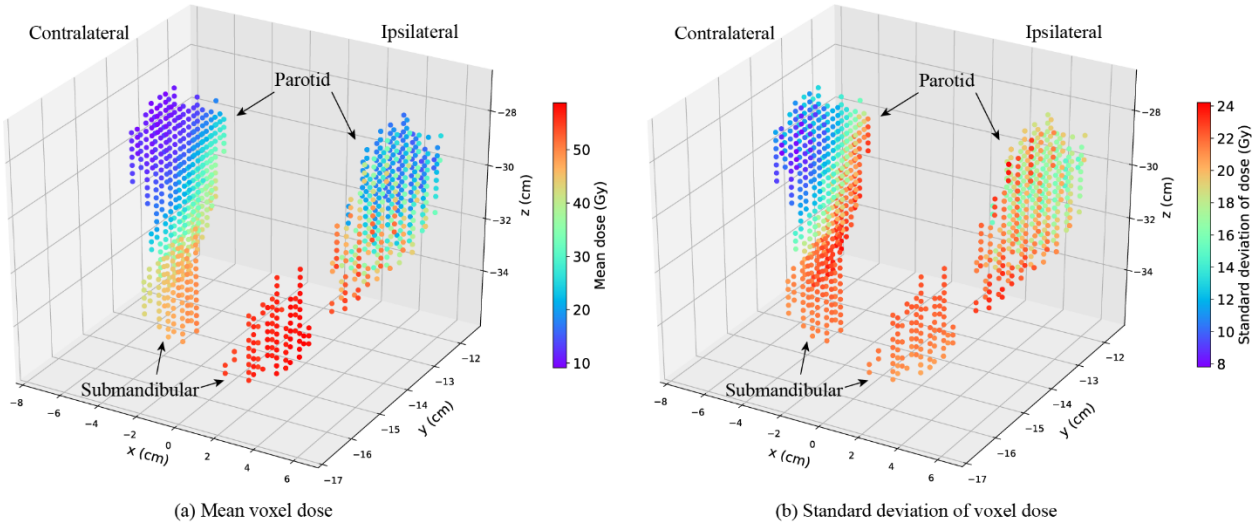
Figure 3.3 The distribution of xerostomia grade at baseline and three months post-RT.

3.3.1.2 Dose Distribution

Figure 3.4 shows the distribution of mean voxel dose and standard deviation of dose in the total 942 voxels of the PG and SMG across the patient cohort. The mean dose in voxels ranges from 9.09 to 58.77 Gy while the standard deviation of dose ranges from 7.81to 24.21 Gy. The ipsilateral SMG has the highest mean voxel doses and the anterior, superior portion of the two PG has the

lowest mean voxel doses. The inferior, posterior portion of the contralateral PG and the two SMG have the highest variation of dose, while the anterior, superior portion of the contralateral PG has the lowest variation of dose in the patient cohort.

The pattern of mean voxel dose generated is consistent with the clinical practice whereby the dose spectrum decreases more superiorly across the PG. The ipsilateral SMG and the tail of the ipsilateral PG typically receive comparable high doses of radiation due to the presence of level II cervical nodal metastases.



(a) Mean voxel dose

(b) Standard deviation of voxel dose

Figure 3.4 The distribution of radiation dose in parotid glands and submandibular glands across the patient cohort.

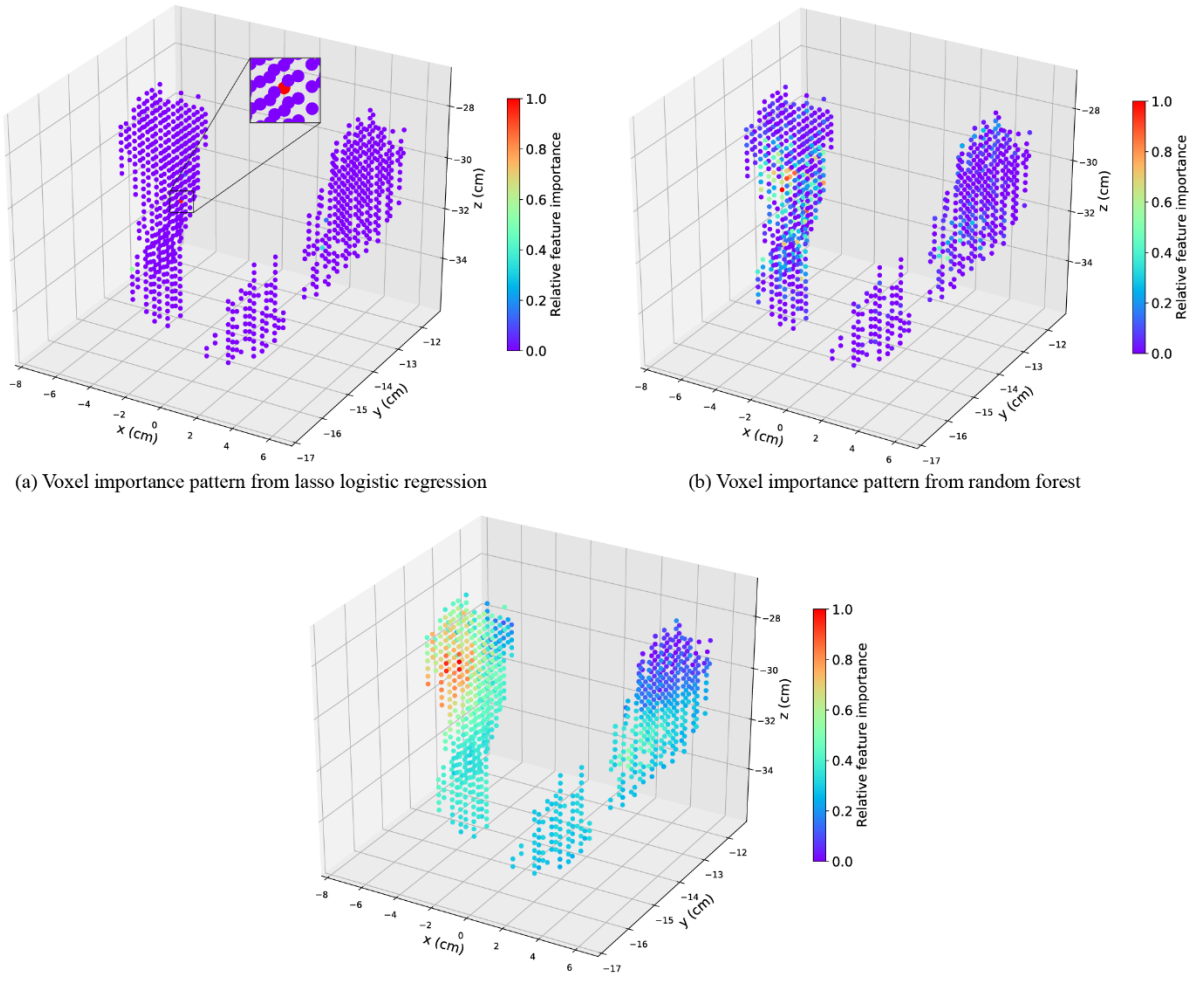
3.3.1.3 Model Performance

After obtaining the radiation dose features, non-dose features, and xerostomia outcomes, the AUC score of the ridge, lasso logistic regression, and random forest were evaluated using 10-fold cross-validation on the obtained dataset. The cross-validation AUC score (out-of-sample score) for ridge, lasso logistic regression, and random forest are: 0.70±0.04, 0.67 ±0.04, 0.69 ±0.06 respectively.

3.3.1.4 Voxel Importance Pattern

Voxel importance patterns for the three methods were shown in Figure 3.5. Ridge and lasso logistic regression models both are well suited for high dimensional data where the number of features is

larger than the number of samples, or many features are correlated [7]. However, given highly correlated voxel-based dose features, lasso yields only one or few voxels from the predictive regions due to ℓ^1-norm regularization. Random forest produces non-stable relative variable importance due to randomly choosing correlated features, but ridge logistic regression produces a stable solution by shrinking weights of correlated features close to each other [7]. The dimension of the voxel radiation dose features is high (over 900 voxels) and much larger than the number of patients (427). Moreover, the radiation dose features are highly correlated for voxels that are spatially close. It's likely that there is a region of voxels, instead of single or few voxels, of which the dose has the largest influence on xerostomia, and other regions have less but not zero influence. The voxel importance patterns in Figure 3.5 from the three methods verified the above reasoning.



(a) Voxel importance pattern from lasso logistic regression

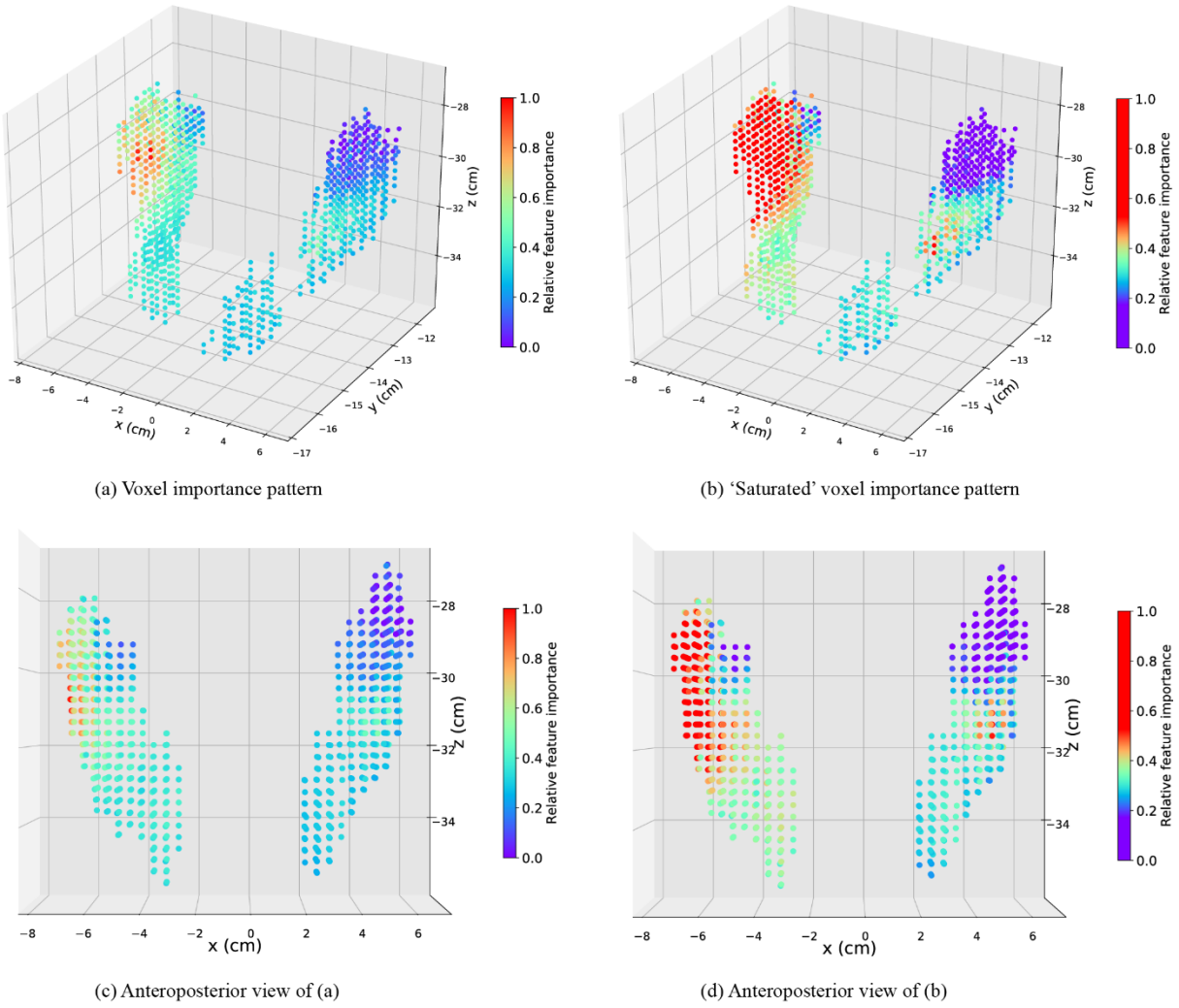
(b) Voxel importance pattern from random forest

(c) Voxel importance pattern from ridge logistic regression

Figure 3.5 Voxel importance patterns learned from the three machine learning algorithms where the color corresponds to the relative importance of each voxel.

Ridge logistic regression also has the best prediction performance. Therefore, ridge logistic regression was used to study the voxel importance pattern, which is shown in Figure 3.6. Red region represents the most important, and violet represents the least important voxels. The dose in different subvolumes in PG and SMG, together all affect xerostomia, but the influence of dose on xerostomia varies across different subvolumes. Specifically, the superior, anterior portion of the contralateral parotid is the most influential region. The medial portion of the ipsilateral PG is also very influential, while the superior portion of the ipsilateral parotid is the least influential region.

The consistency of the voxel importance pattern obtained from ridge logistic regression was tested on five different random samples. The lowest Pearson correlation coefficient among the weights learned from the five random samples is 0.85, which indicates that the voxel importance pattern is fairly consistent given sample randomness.



(a) Voxel importance pattern

(b) 'Saturated' voxel importance pattern

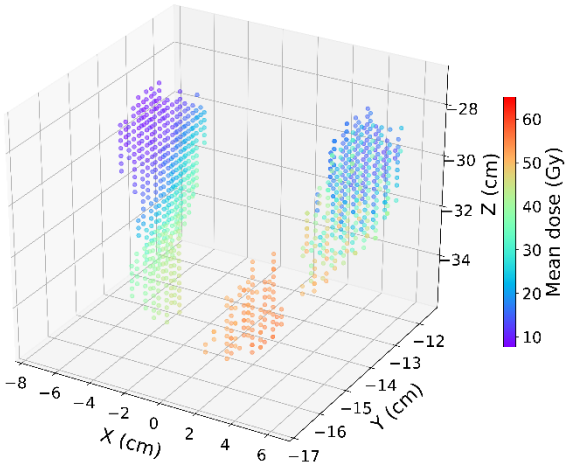
(c) Anteroposterior view of (a)

(d) Anteroposterior view of (b)

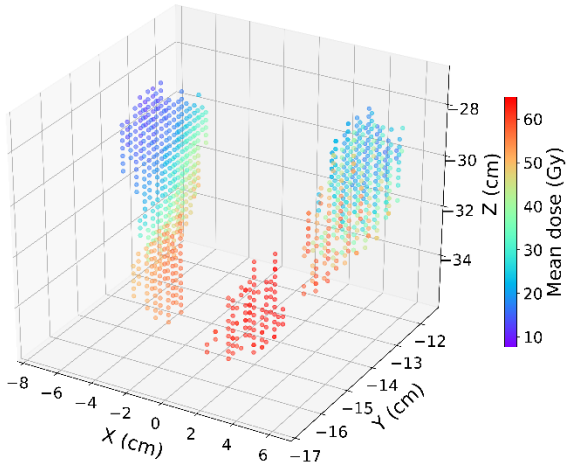
Figure 3.6 Voxel importance pattern from ridge logistic regression. (b): a different visualization of the same voxel importance result where voxel importance values that are one standard deviation away from the mean were "saturated" to increase the resolution of voxel importance closer to the mean value of the voxel importance. (c): anteroposterior view of (a). (d): anteroposterior of (b).

3.3.1.5 Dose Comparison using Statistical Test

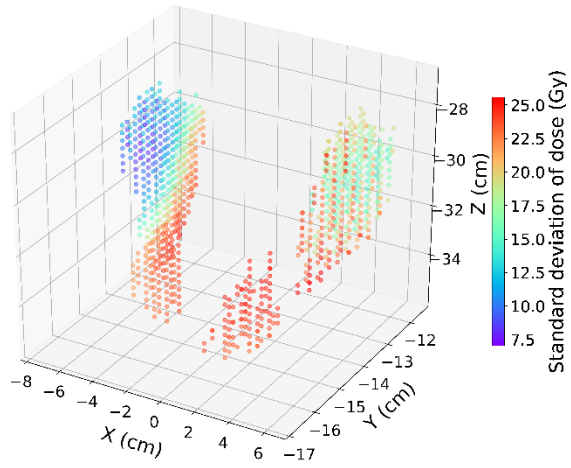
We further compared the dose distribution between the two xerostomia groups. The goal is to directly see how different distribution of radiation dose was delivered to the two xerostomia groups. First, we visualized the dose distribution of each group and the dose difference between them. Second, performing permutation test which is a non-parametric two-sample statistical hypothesis test. Figure 3.7 shows the mean dose and standard deviation of dose for the two patients group categorized as acute xerostomia group versus non-acute xerostomia group. It shows that the mean dose in the submandibular glands and inferior part of contralateral PG is higher for the patients who developed acute xerostomia, while the patients who didn't develop acute xerostomia have a higher variation of dose among them in the submandibular glands, ipsilateral PG, and inferior part of contralateral PG. The dose difference suggests that the acute xerostomia group patients overall were consistently treated with a higher dose in the parotid and submandibular glands.



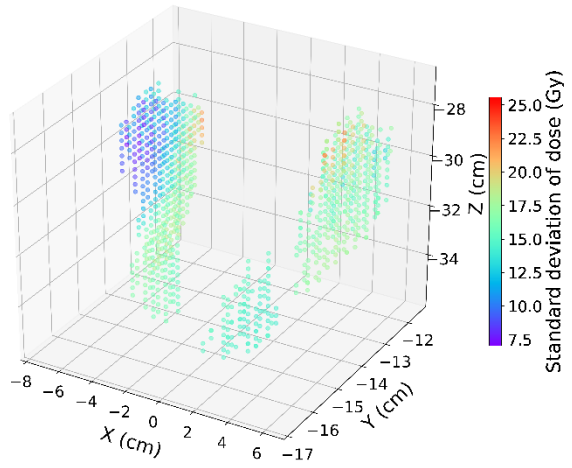
(a) Mean dose, acute xerostomia: no



(b) Mean dose, acute xerostomia: yes



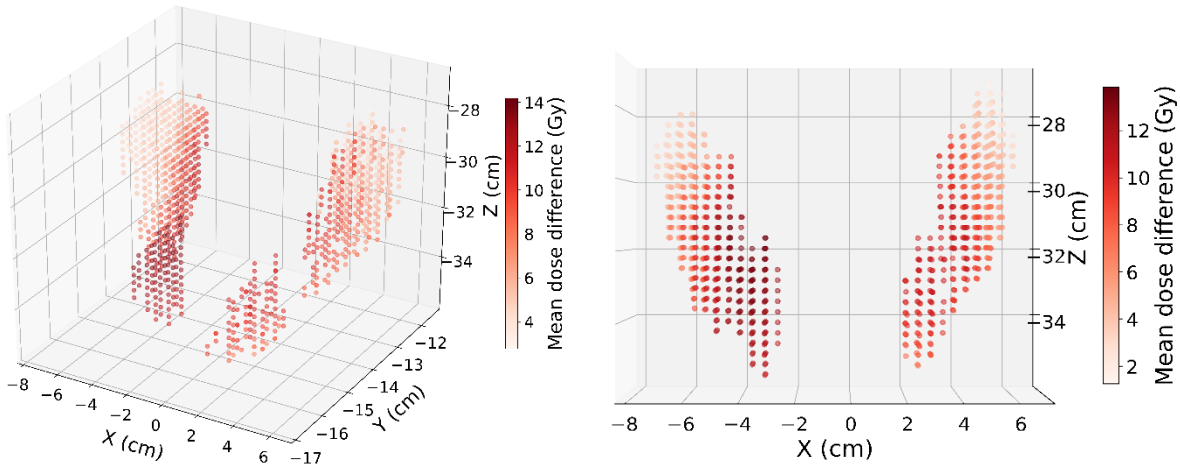
(c) Dose variation, acute xerostomia: no



(d) Dose variation, acute xerostomia: yes

Figure 3.7 Dose distribution for patients group who developed acute xerostomia versus who didn't.

To see the dose difference more clearly, we plotted the mean dose difference between the two groups as the mean dose of acute xerostomia group minus the mean dose of non-acute xerostomia group shown in Figure 3.8. It shows that the acute xerostomia group, on average, have higher dose across the entire parotid and submandibular glands on both ipsilateral and contralateral side. A particular higher dose was delivered to the contralateral submandibular gland and inferior part of the contralateral PG for the acute xerostomia group.



(a) Side view (b) Anteroposterior view

Figure 3.8 Mean dose difference as mean dose of acute xerostomia group minus mean dose of non-acute xerostomia group.

To quantitatively compare the dose difference between the two groups. We applied permutation test, a nonparametric statistical significance test for comparing two samples [65]. The

advantage of permutation test is that, as a nonparametric test, it doesn't require assumptions on the sample distributions, which are often required by parametric tests such as a two-sample t-test. Ideally, we need to compute the test statistics for all the permutation samples, which is often infeasible for large sample size. A large number of random permutations is often performed to obtain the distribution of the test statistics under the null hypothesis. For our study, we randomly permutated the samples 10000 times, which yields a significance level as small as 0.0001. The null hypothesis for the permutation test is that the radiation dose of the acute xerostomia group and the non-acute xerostomia have the same distribution. Thus their mean dose should be the same. Here, we performed a one-sided hypothesis test and set the alternative hypothesis as $\mu_{acute\ xero} > \mu_{non-acute\ xero}$, where $\mu_{acute\ xero}$ and $\mu_{non-acute\ xero}$ are respectively the mean dose for the acute xerostomia group and non-acute xerostomia group. The permutation test was performed for comparing the mean dose level in each voxel between two groups. The steps for the permutation test are:

1. Compute the actual sample test $t = \mu_{acute\ xero} - \mu_{non-acute\ xero}$
2. Put the radiation dose in a certain voxel from two patient groups together in an array.
3. Randomly permute the dose array while keeping the patient group label the fixed.
4. Compute the mean dose difference $T = \mu_{acute\ xero} - \mu_{non-acute\ xero}$, between two patient groups using the permutated dose array.

5. Repeat steps 2-3 for a large number of times (10000 in our case) and obtains the distribution of the test statistics $T_i = \mu_{acute\ xero_i} - \mu_{non-acute\ xero_i}$, where $i = \{1,2,\dots,10000\}$.

6. Compute the *p*-value as the percentage of permutated samples whose test statistics are larger than the actual sample test statistic, i.e. $\frac{\sum_i I(T_i > t)}{10000}$, where I is the indicator function.

Figure 3.9 shows the distribution of test statistics and its actual sample test statistic indicated by the red line as results of the permutation test for comparing dose in one of the voxels. *p*-value is the area of the histogram on the right side of the red line.

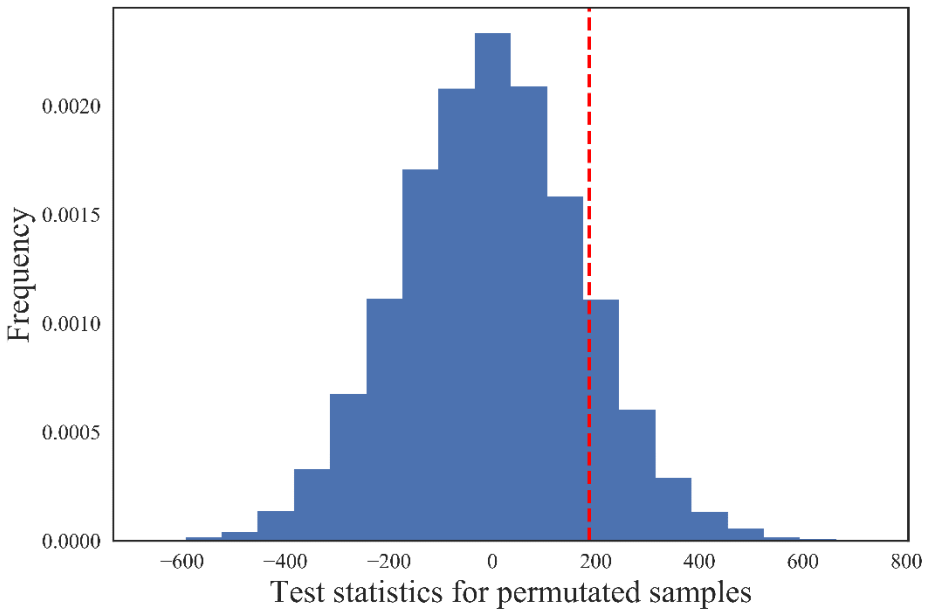
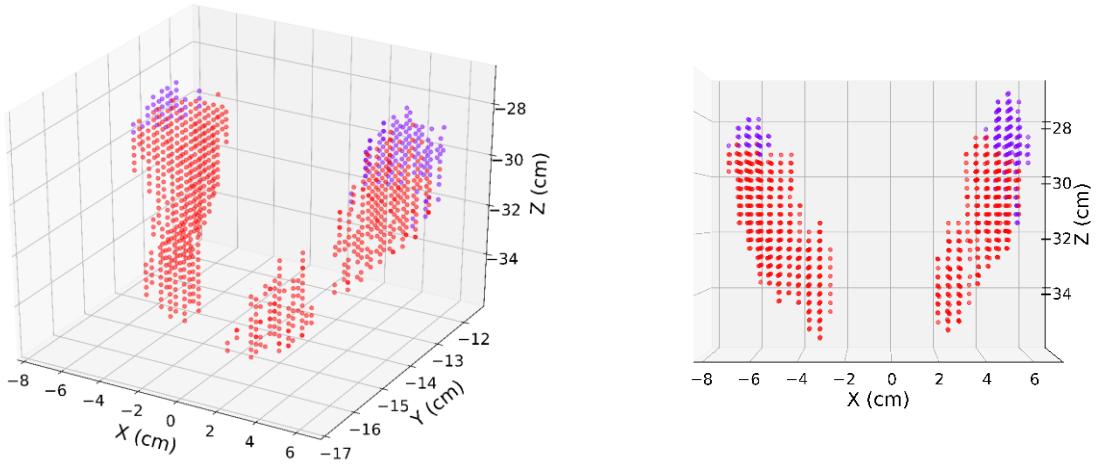


Figure 3.9 Distribution of test statistics using permutation test and the actual sample test statistic for comparing mean dose in two patient groups in one of the voxels.

After performing permutation test for dose in each voxel, we obtained the *p*-values and showed them in Figure 3.10. The voxels in red are the regions where the null hypothesis was rejected, and the mean dose is statistically significantly different. Subplots (a) and (b) show that the mean dose across the red regions in the parotids and submandibular glands in the acute xerostomia patient group are statistically significantly higher than the non-acute xerostomia patient

group with a significance level of 0.0002, but we can't reject the null hypothesis that the mean dose levels are the same for the two patient groups in the superior portion of PGs (blue region).



(a) (b)

Figure 3.10 Distribution of 1 minus the *p*-values for comparing the mean dose of acute versus non-acute xerostomia patient groups. (a): regions where *p*-values above or equal to 0.0002 were highlighted in blue, and p-values less than 0.0002 were highlighted in red. (b): anteroposterior view of (a).

3.3.2 Xerostomia Recovery

After studying the dose influence pattern on acute xerostomia which was caused by radiation injury, we performed the same analysis on xerostomia recovery and sought to compare the results between acute xerostomia and xerostomia recovery to obtain a better understanding of the dose

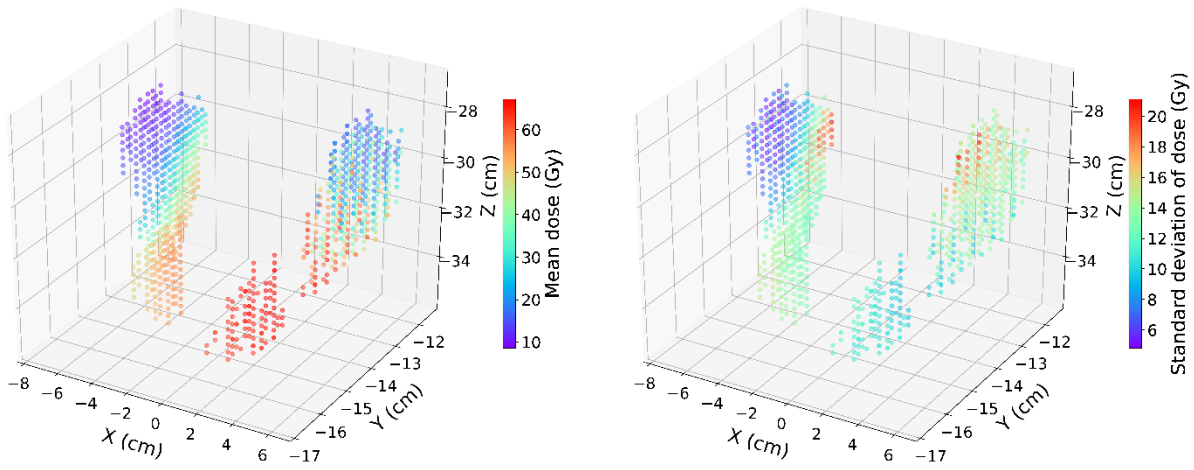
effect on xerostomia. We expected the dose influence pattern for xerostomia recovery to be different from acute xerostomia.

3.3.2.1 Xerostomia Recovery Outcome

To define xerostomia recovery, we took xerostomia grade measurements from three time periods: baseline (before or within the first week of treatment start), between the end of the radiotherapy and 18 months after radiotherapy, and beyond 18 months after radiotherapy. We took the maximal xerostomia grades between the end of and 18 months after radiotherapy as the xerostomia measure for the second period. Similarly, we took the maximal xerostomia grades beyond 18 months after radiotherapy as the xerostomia measure for the third period. Severe xerostomia was defined as xerostomia grade equal to or above two. The patient cohort consists of patients who do not have xerostomia grade at baseline and have xerostomia measurements available beyond 18 months after radiotherapy. Patients who did not recover from xerostomia was defined as whose xerostomia pattern over the three time periods is $0\backslash 1\backslash 1$, where 0 represents no severe xerostomia and 1 represents severe xerostomia. The xerostomia pattern for recovered patients is $0\backslash 1\backslash 0$. We treated the xerostomia recovery prediction problem as a binary classification problem as well.

The total number of patients in this recovery study cohort is 146 (non-recovered/recovered: 32/114). The same set of voxel dose features and non-dose features as the initial acute xerostomia study was used.

3.3.2.2 Dose Distribution



(a) Mean dose (b) Standard deviation of dose

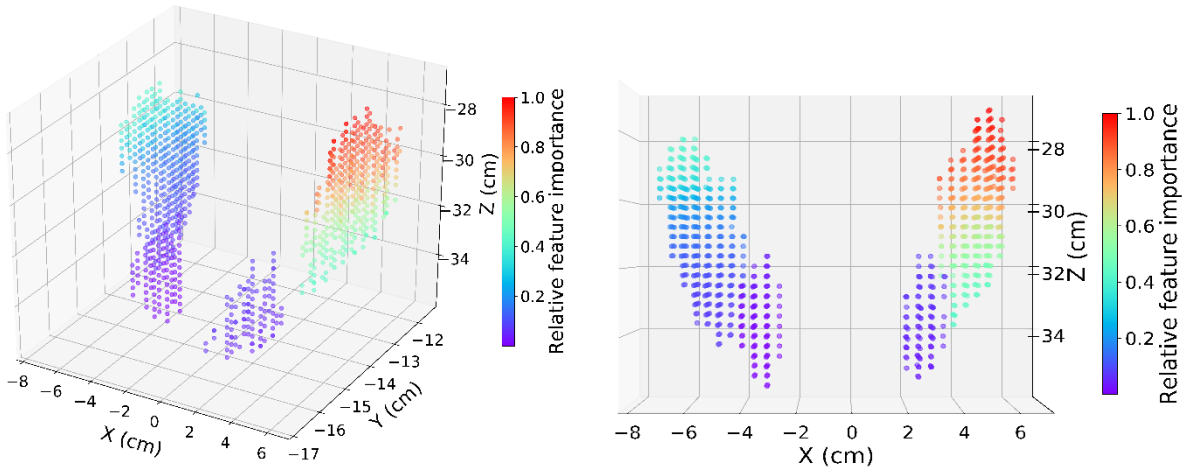
Figure 3.11 The distribution of radiation dose in parotid glands and submandibular glands across the xerostomia recovery patient cohort.

Figure 3.11 shows the distribution of mean voxel dose and standard deviation of voxel dose in the PG and SMG across the xerostomia recovery patient cohort. The mean dose in voxels ranges from 8.66 to 67.19 Gy while the standard deviation of dose ranges from 4.84 to 21.22 Gy. The ipsilateral SMG has the highest mean voxel doses and the anterior, superior portion of the two PG has the lowest mean voxel doses. The inferior, posterior portion of the contralateral PG and the superior portion of ipsilateral PG have the highest variation of dose, while the anterior, superior portion of the contralateral PG has the lowest variation of dose in the patient cohort.

The distribution of mean dose for recovery cohort is very similar with acute xerostomia cohort, while the acute xerostomia cohort has a much larger variation of dose likely because its sample size is three times of the recovery cohort.

3.3.2.3 Voxel Importance Pattern

After applying ridge logistic regression model on the xerostomia recovery data, we plotted the voxel importance pattern in Figure 3.12. We indeed found that the relative voxel importance pattern for xerostomia recovery is different from the pattern for acute xerostomia. The voxel importance pattern shows that the dose level in ipsilateral PG is most influential on xerostomia recovery, particularly the superior portion of the ipsilateral PG.

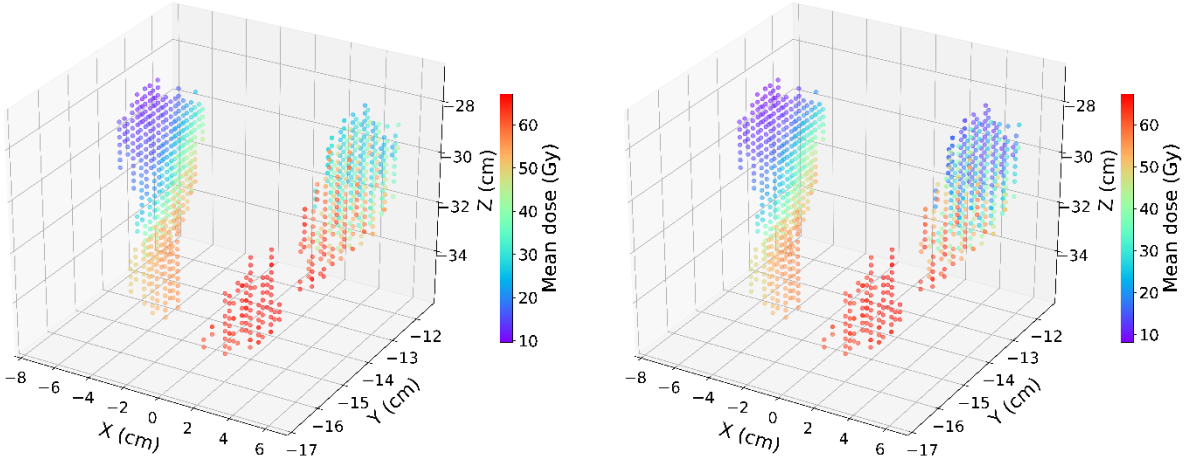


(a) Side view (b) Anteroposterior view

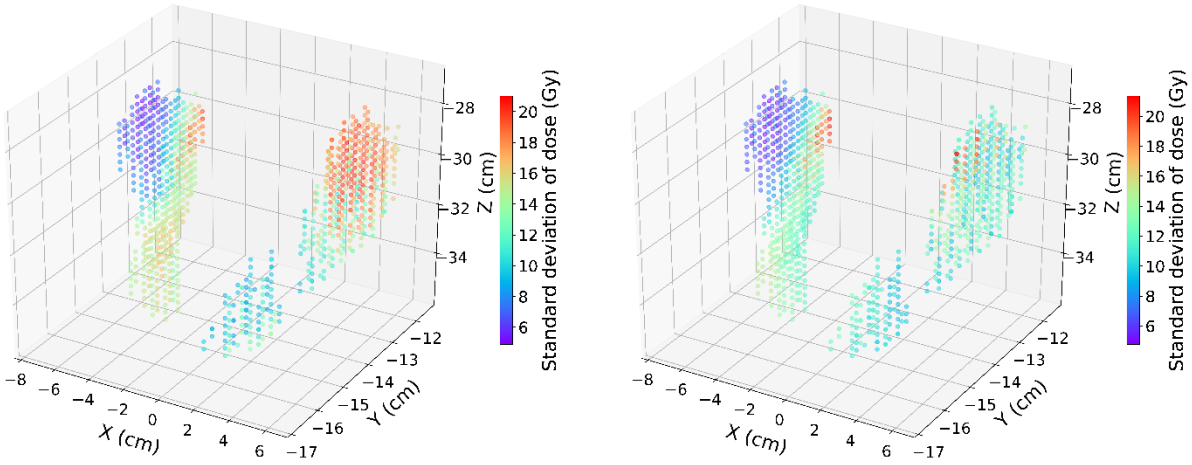
Figure 3.12 Voxel importance pattern from ridge logistic regression for xerostomia recovery.

3.3.2.4 Dose Comparison using Statistical Test

To see if the voxel importance pattern is related to the dose distribution in the two different recovery groups, we compared the dose distribution between the two groups by visualizing the mean dose and standard deviation of voxel dose for each cohort in Figure 3.13. Subplots (a) and (b) show that the non-recovered patients were treated with a higher dose in the superior portion of ipsilateral PG on average, while the dose level other regions in the PG and SMG are about the same. The non-recovered patients also clearly have higher dose variation in the superior portion of ipsilateral PG, which indicates that the recovered patients received a consistently lower dose in the superior portion of ipsilateral PG.



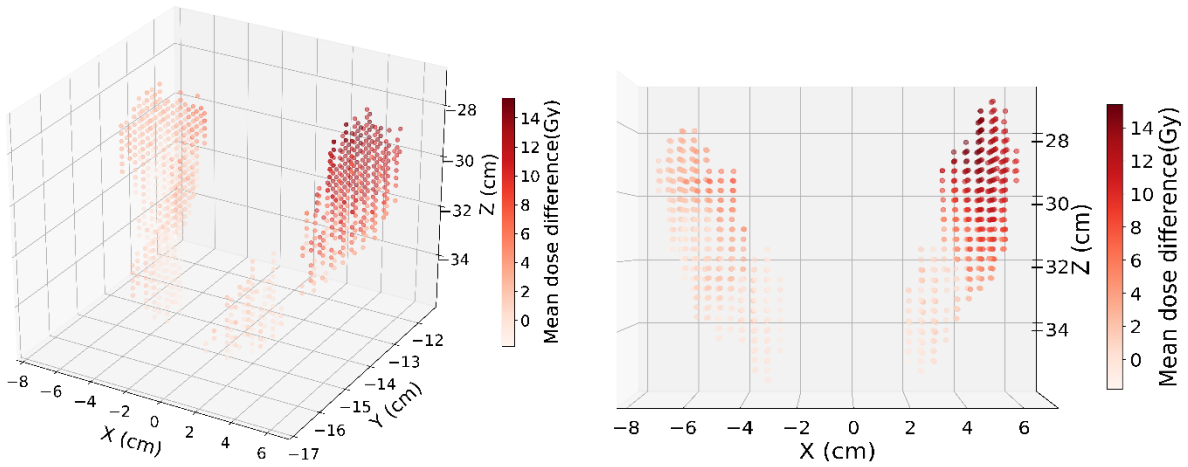
(a) Mean dose for not recovered group (b) Mean dose for recovered group



(c) Dose variation for the not recovered group (d) Dose variation for the recovered group

Figure 3.13 Dose distribution for not recovered patients group versus recovered group.

To more directly see the dose difference between the two recovery groups, we plotted the mean dose difference as the mean dose of not recovered group minus the mean dose of the recovered group shown in Figure 3.14. It shows that the superior portion of the ipsilateral PG of the not recovered group received a much higher dose, while dose in other regions is about the same. The dose difference indicates that being able to recover xerostomia is associated with receiving a lower radiation dose in the ipsilateral PG, especially the superior portion.



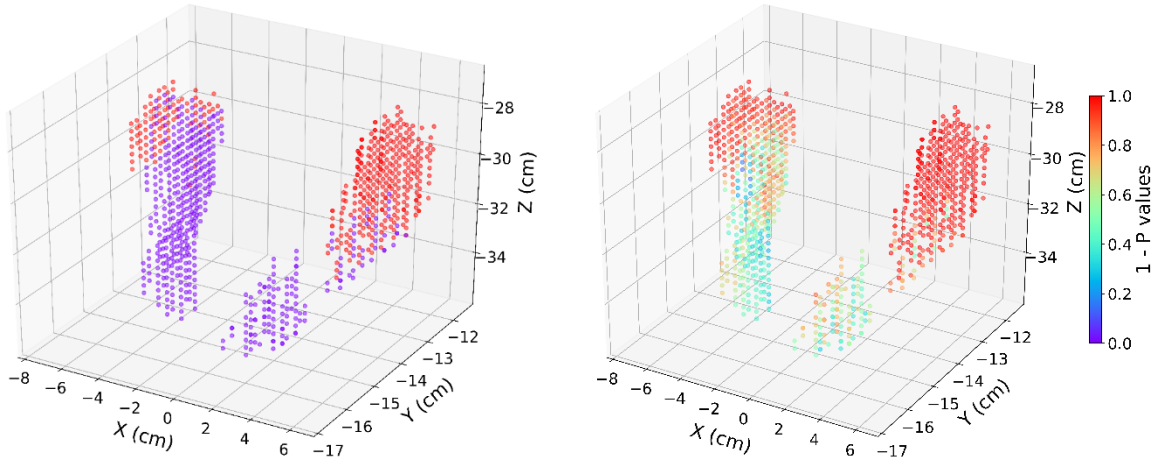
(a) Side view (b) Anteroposterior view

Figure 3.14 Mean dose difference as mean dose of the not recovered group minus mean dose of the recovered group.

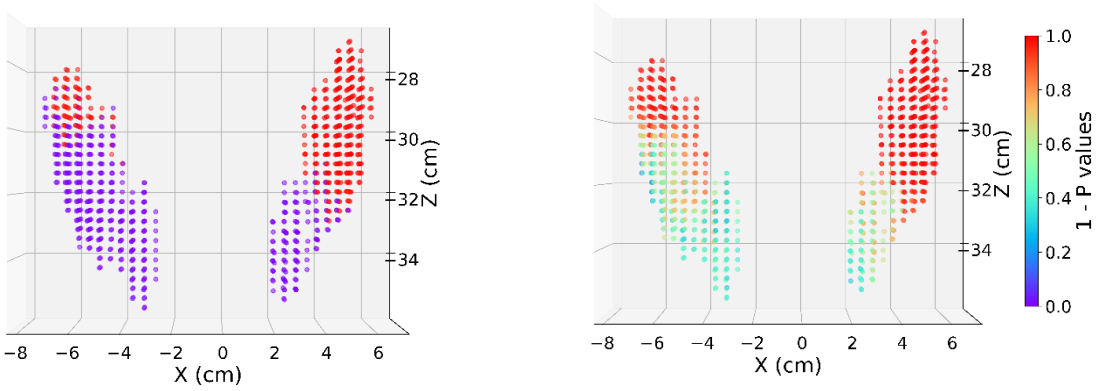
To quantitatively check if the dose level is statistically significantly different in any voxels between the two recovery groups, we performed the one-sided permutation test on all the voxel dose for the recovery cohort. The distribution of p-values was shown in Figure 3.15, and it shows that the mean dose in ipsilateral PG and superior portion of the contralateral PG is significantly different, particularly the dose is higher for not recovered patients, at a significance level of 0.05. In other words, we rejected the null hypothesis that the mean dose levels are the same between the two groups in favor of the alternative hypothesis, $\mu_{not\ recovered} > \mu_{recovered}$.

This result matches with the voxel importance pattern obtained using ridge logistic regression on xerostomia recovery. They both indicate that higher dose in the ipsilateral PG, particularly the superior portion, is associated with higher probability of not being able to recover from xerostomia.

The voxel importance pattern for recovery is more symmetric than the pattern for acute xerostomia. However, both voxel importance patterns show that dose in PGs, especially the superior portion is most influential on xerostomia, while the SMGs are relatively not influential. We will discuss in detail what the different voxel importance patterns implicate and how they might be related to the specific dose distribution in our cohorts.



(a) (b)



(a) (b)

Figure 3.15 Distribution of 1 minus the *p*-values for comparing the mean dose of not recovered versus recovered from xerostomia patient groups. (a): regions where *p*-values above 0.05 were highlighted in blue. (b): distribution of 1 minus *p*-values. (c): anteroposterior view of (a). (d): anteroposterior view of (b).

3.4 Discussion

The results presented demonstrate the successful application of the use of spatially explicit radiation dose features in predicting xerostomia in HNC patients, leading to the identification of important parotid dose subvolumes associated with the risk of severe xerostomia at three months following radiotherapy. To the best of our knowledge, this analysis used the largest xerostomia HNC patient cohort in literature, which enables robust modeling results and conclusions. The analysis confirmed human evidence of parotid subvolumes being at a greater risk of contributing

to radiation-induced xerostomia as initially proposed in humans by van Luijk et al. [57]. Modeling the radiation dose spatially explicitly provides insights into the spatial dependence of dose in contrast to the use of DVH features alone. For instance, we are aware of one existing study that used a similar voxel-based dose analysis for acute dysphagia in HNC patients [66]. The authors applied a statistical test to compare the voxel-based dose distribution between patients with grade >= 3 versus with grade <3 dysphagia. In contrast, we combined voxel-based dose modeling with supervised machine learning algorithms to build a predictive model for xerostomia and also learned how spatial dose influences xerostomia, highlighting the flexibility of this methodology. This also demonstrates that our methodology can be applied to other radiotherapy-related head and neck toxicities such as dysphagia and trismus.

The distribution of average dose shows that the most important region found for acute xerostomia, i.e., the superior, anterior portion of the contralateral parotid, is within the low dose region in the patient cohort. This is consistent with results from our prior studies using DVH features which demonstrated the impact of a low dose bath to both PG on the risk of grade >=2 xerostomia [59]. We believe if this region is treated with high radiation dose, it's likely that the other regions were treated with even higher dose, leading to a much higher risk of developing xerostomia. Further, we would expect the superior portion of ipsilateral PG also to be more predictive as it's also a low dose region. However, the voxel importance pattern result is not showing this. Rather the medial portion of the ipsilateral PG is more predictive. We hypothesize

that in the ipsilateral PG, the medial portion is more important because it's very close to the ductal region. Radiation damage to this medial portion will render dose in the superior portion much less important in predicting xerostomia. This matches the hypothesis that we may have damaged the duct responsible for transporting saliva for the patients that received a high dose in that region.

The voxel importance pattern for xerostomia recovery shows that the dose level in ipsilateral PG is most influential on xerostomia recovery, particularly the superior portion of the ipsilateral PG. The superior portion of contralateral PG is relatively important on the contralateral side. The superior portion of the ipsilateral PG is also the low dose bath region in the recovery cohort. However, the machine learning algorithm is not simply identifying the low dose bath region as the most influential region as the support portion of the contralateral PG has the lowest mean dose but the algorithm didn't identify that region as most influential.

For both voxel importance patterns, the machine learning algorithm didn't exactly identify the low dose bath region as the most influential region regarding dose effect on xerostomia. Neither of the voxel importance patterns is symmetric, which is likely caused by the non-symmetric dose distribution delivered across the ipsilateral and contralateral PGs. Combing the two voxel importance patterns, we can see it's clear that dose in the PGs is more influential than the dose in the SMGs for our cohort. We hypothesize that this can be caused by two reasons. First, the dose in SMGs is influential on xerostomia. The reason that the algorithm didn't identify SMGs as influential regions is that, in our cohort, the SMGs were consistently treated with very high dose,

especially the ipsilateral side. Even though for acute xerostomia study, acute xerostomia patients received statistically significant higher dose than non-acute xerostomia patients and there is enough variation of dose in SMGs, we believe that the dose variation won't have a differential effect on xerostomia outcome if the dose is very high and exceeds a certain threshold. In other words, we expect the dose effect on xerostomia to be nonlinear and is diminishing when it exceeds a certain high threshold. Second, the dose in the SMGs may indeed have a very limited effect on xerostomia outcome.

On the other hand, the PGs received a relatively lower dose on average for our cohort, and the dose variation in PGs was shown to have a larger differential effect on xerostomia outcome. This was well explained, particularly, by the xerostomia recovery analysis. The superior portion of ipsilateral PG was identified as the most influential region by the machine learning algorithm. The dose distribution in the two recovery groups shows that region has the largest mean dose difference as high as 14 Gy, while all the other regions have about zero mean dose difference. This region is also in the low dose bath region on the ipsilateral side. The pattern of mean dose difference of the xerostomia recover cohort is much simpler than the pattern of the acute xerostomia cohort, and it makes it much easier to see how the influential region identified by the machine learning algorithm is related with the dose distribution.

Overall, the results from acute xerostomia and xerostomia recovery show that the dose in the superior portion of the two parotid glands (low dose bath region) are relatively most influential

on developing xerostomia. Besides, higher dose delivered across the PGs and SMGs are associated with more severe xerostomia (both acute and long-term). However, the findings are specific to our cohort and affected by the limited dose variation in the cohort. This highlighted the challenge that it's difficult to identify the causal effect of level of the dose delivered to organ subvolumes on xerostomia using observational study. It also highlighted the complexity of the dose-effect and potential spatial dependency of dose-effect across organ subvolumes. For instance, we would expect the dose effect on xerostomia in the PGs to be more or less symmetric, but our results didn't reveal this in each xerostomia analysis, which is, again, likely caused by the specific dose distribution for each study cohort.

The findings provide insight into the importance of bilateral parotid injury and underscore the importance of carefully determining the clinical indications for bilateral cervical nodal irradiation along with a careful delineation of how superior and lateral the cervical nodal planning target volumes encroach upon the medial and superior aspects of the PG. Moreover, the analysis also demonstrated to a lesser degree, the importance of the subvolumes in the contralateral SMG contributing to the severity of the xerostomia at three months post-RT. This highlights the clinical implications and importance of reducing the volume of cervical neck irradiation as a clinical strategy to de-intensify the current chemoradiation treatment paradigms for head and neck squamous cell carcinomas, especially for HPV-associated oropharyngeal carcinomas.

Several additional limitations of our analysis need to be recognized. First, the study population consists of patients treated in a single local hospital. The specific voxel importance pattern obtained with the algorithmic modeling may be valid only for specific patients and specific dose distribution planned at this hospital. Second, for acute xerostomia prediction outcomes, we used last-observation-carried-forward to obtain the data for patients who dropped out before three months post-RT. Our longitudinal xerostomia outcomes data show that there is minor xerostomia recovery from the end of treatment to three months post-RT. Therefore, our prediction outcome data could be slightly biased towards a more severe xerostomia grade. Third, the pattern of dose effect on xerostomia we learned in this study represents only association or correlation between radiation dose and xerostomia. As most patients have similar patterns of dose, it is challenging to evaluate dose-response outside of the range of patterns delivered to patients in the database. No conclusion on the causal effect between dose and xerostomia was established in this observational study. Finally, this study only included radiation dose in PG and SMG, while there is a study reported the mean dose to oral cavity is associated with xerostomia as well [51].

For future studies, we are going to apply this approach to HNC patients in other hospitals, which have different patient characteristics and radiation treatment plans to validate our approach and the voxel importance pattern we learned. Ultimately, we want to learn causal effect between radiation dose and xerostomia, which is difficult using observational studies. Either experimental study or advanced causal inference analysis should be conducted. Our team recently proposed a

causal inference technique, i.e., causal sufficient dimension reduction, for high dimensional treatments problems [67], which we are currently applying to our problem. Finally, we will also include dose in the oral cavity in our study to investigate how dose in subvolumes in the oral cavity is associated with xerostomia.

3.5 Conclusion

We have identified that dose to the specific subvolume in the PG and SMG, i.e., the superior, anterior portion of the contralateral PG is most predictive of acute xerostomia and superior portion of ipsilateral PG is most predictive of xerostomia recovery. Those influential regions all lie within the low dose bath region in our study population. We also found the medial portion of the ipsilateral PG to be predictive of acute xerostomia. We believe our methodology and the local dose effect pattern identified can help improve radiation treatment planning and reduce xerostomia for HNC patients.

Chapter 4

Wood Chip Trade Response to Renewable Energy Policies

4.1 Introduction

The U.S. Renewable Fuel Standard (RFS) mandates blending of biofuel in road gasoline and diesel [68]. The mandated volumes of biofuel are scheduled to increase every year to 2022, and cellulosic biofuel made from non-food feedstocks is the fastest growing component of this mandate. Although actual production of cellulosic biofuel has consistently fallen below mandated volumes in every year of the RFS program, commercial-scale cellulosic biofuel facilities continue to be built utilizing a wide range of feedstocks. Environmental Protection Agency (EPA) has approved some feedstocks that can be used to produce cellulosic biofuel that is eligible for RFS support,

including perennial grasses like miscanthus and switchgrass, agricultural residues like corn stover, and forestry residues [68]. The agency is considering expanding this list to include pulpwood from whole trees [69], which, if approved, could lead to increased demand for wood harvests. U.S. consumption of wood for electricity has been rising since 2009 [70]. At the same time, the EU's Renewable Energy Directive is driving very high imports and consumption of wood pellets [71], which are produced from wood chip and other pulpwood product. The EU has been increasingly dependent on biomass import for energy. Its total imports of wood chip both for energy and pulp production have increased from 6 million cubic meters in 1997 to 16 million cubic meters in 2011 according to the FORESTAT database [72]. Wood pellets were identified to be efficient to be used for co-firing to generate renewable electricity in Germany and Austria [73]. However, co-firing coal with wood pellets is currently not economically feasible within the U.S. due to the recent U.S. natural-gas boom [74].

With these two policies working together to drive up demand for wood harvests, where will the additional supply come from? The answer is important to understanding the lifecycle environmental impacts of the potential pulpwood to biofuel pathway under the RFS. Some additional supply would very likely come from the Southeastern U.S. where most American pulpwood is produced. However, due to the EU's phytosanitary measures [75], US's export of softwood chip to the EU has been very limited because of the existence of nematodes in US's softwood chip. Therefore, the EU's increasing demand would need to be satisfied by other regions

instead. If the additional wood is produced from new forestry plantations on previously unforested land, cellulosic biofuel could deliver significant carbon savings. If on the other hand, the increased wood supply comes from existing forests in countries with weak forest protection policies and enforcement, a pulpwood biofuel pathway in the U.S. would likely cause a net increase in emissions compared to fossil gasoline or diesel [76].

Most previous studies about wood bioenergy markets were conducted at the country level, specifically for European countries such as Austria, Norway, Italy, and Poland. Those studies in general focus on three aspects: first, bioenergy potential of a certain region; second, demand, supply and production of certain bioenergy; and third, bioenergy usage in a certain country. For example, Nilssona et al. analyzed the status and potential of bioenergy in Poland in 2006 and found that firewood for heating was the main bioenergy usage [77], which consisted of 95% of renewable energy usage in 2003. Nilssona et al. concluded that Poland's bioenergy market and policy were undeveloped even though it had a large potential for bioenergy [77]. Paiano et al. estimated the bioenergy potential in Italy and found 2.7% of the gross Italian energy consumption in 2013 could be generated from residual biomass, which could save about 52 Mt CO_2eq emission for Italy per year [78]. Trømborg et al. analyzed the effect of various bioenergy policies on the usage of forest-based bioenergy in Norway using a spatial partial equilibrium model and found the share of bioenergy in the Norwegian energy market was much lower than other EU countries due to low electricity price and lack of heating facilities [79]. They concluded that policy incentives including

subsidies, deposit grant, and feed-in systems could significantly increase Norway's bioenergy production. Trømborg et al. also gave a detailed presentation of the forest biomass potentials for heating in Norway in 2011 and concluded it is unlikely the government target of 14 Twh more bioenergy by 2020 can be met [80]. Also, those studies mainly focused on wood pellets but not wood chip.

Few studies have been done about the international trade of wood bioenergy, especially for wood chip. One study presented an overview of the historical international trade flow, bioenergy policies and market factors for solid biofuel such as wood pellet, wood chip, and roundwood in main markets including the EU, North America, Russian Federation, and Japan [81]. They identified that wood pellets had become the most traded solid biofuel as a globally traded commodity and its trade increased from 8.5 PJ to 120 PJ from 2000 to 2010. Another study reviewed the market factors and policies for global wood pellet market and presented the opportunities and challenges for the wood pellets industry [82]. They expected that the EU would remain the main wood pellets market and East Asia's market would be further expanded. The only study on global wood chip trade for energy was done by Lamers et al. [75]. They presented the historical global trade data of wood chip and estimated that energy-related wood chip trade volume was less than 10% annually. Also, they identified that the key constraint of trading wood chip for energy is production and transportation cost. These studies provided an extensive overview of the

wood chip trade data. Building on this data, our study takes a further step and analyzes the wood chip trade changes under different EU and U.S. renewable energy policy scenarios.

Junginger et al. also identified logistics including transportation as the major barrier for solid biomass commodities due to their low energy density and a relatively low value [83]. A recent financial analysis of the transport of wood chip from the USA to Germany estimated that transporting wood chip per weight is more than twice as expensive as transporting wood pellets because wood pellets have a much higher density [84].

In addition, phytosanitary requirements are another barrier to the trade of wood chip that is infected by insect pests. For example, export of softwood chip from the US to the EU was restricted due to the EU's phytosanitary requirement regarding wood chip as previously mentioned.

The main contribution of this study is we build the first global trade model for wood chip and analyze how local energy policy from U.S. and EU will affect the global market for wood chip. Specifically, we find that wood chip exports from tropical regions would increase significantly. Implementation of sustainability criteria for biomass should focus on these regions. To ensure the imported biomass feedstock is sustainable, EU has initiated the BioTrade2020plus project. Iriarte et al. suggested the sustainability criteria and assessed the sustainability risks for biomass including wood chip focusing on current and potential future major sourcing regions including Latin America, Asia, and Russia [85]. We estimate supply elasticity and transportation

cost of wood chip using positive mathematical programming, an automatic calibration technique that has been extensively used to overcome limited availability of trade data and supply data.

Also, our analysis can be a springboard towards deeper analysis by including other policy, environmental and technical factors such as the implementation of sustainability criteria, technology changes, and forest growth.

4.2 Methods and Materials

4.2.1 Data Preprocessing

The wood chip trade data were downloaded from the FORESTAT database [72]. The commodity extracted from this database was "wood chip and particles." The data included the quantity and value of wood chip traded between countries. Quantity was measured in cubic meters and value was measured in thousands of U.S. dollars. Export values are reported as free on board value while import values consist of cost, insurance, and freight, according to the Food and Agricultural Organization of the United Nations (UN FAO) [86].

It's important to note that the wood chip trade data we used is the entire trade of wood chip, not solely the bioenergy-related wood chip trade. We didn't separate the wood chip trade data based on its end-use, i.e., used for bioenergy or paper production, for three reasons. First, our goal is to analyze the changes of wood chip trade due to renewable energy policies. Our goal is not tracking the bioenergy-related trade of wood chip. Otherwise, we indeed need to separate wood

chip trade based on its end-use. For example, in 2006, Hillring analyzed the trade patterns for forest product and wood fuel [87]. For wood fuel, Hillring's analysis focused on charcoal trade. Wood chip trade data was presented in the analysis but not categorized based on its end-use. Later in 2012, Lamers et al. estimated that annually reported energy-related wood chip trade volumes were less than 10% based on anecdotal evidence, literature review and personal assumption, for example, assuming global trade of wood chip for energy was exclusively towards the EU [81]. We believe wood chip traded for paper production (pulpwood) can also be used for bioenergy. For example, as Olsson and Hillring mentioned, in 2009, the global financial crisis reduced demand for pulpwood from Swedish pulp & paper producers, which led to the excessive export of pulpwood for energy to Denmark [88]. Therefore, we modeled the entire demand and supply of wood chip, but our scenarios only modeled changes in policies from energy and changes in energy policy result in absolute changes in wood chip consumption globally. Second, there is no trade data currently available for wood chips used for bioenergy in international statistics because current six digits international trade code for wood chip doesn't differentiate by end-use. Third, indirect trade of woody biomass makes it further complicated to separated energy-related wood chip trade from wood chip traded for pulp and paper production. For instance, for the Kraft pulp production, part of the pulpwood is used to produce heat in the pulp mills [89][90]. This part of pulpwood was not explicitly traded for energy but still ends up in energy production. Therefore, we used the overall wood chip trade data for our analysis given the goal of our study.

The dataset contains both volume and prices for wood chip between countries. This country level dataset provides us the flexibility to do analysis both at the country level or aggregated regional level. However, there were some discrepancies in the raw data from year to year. For example, the main discrepancy in quantities was that imports and exports between countries did not match. Various reasons exist for trade data discrepancies. For example, country A's exports could arrive at country B the following year, leading to total exports not matching total imports for the year. Moreover, some exporters may underreport to reduce tariff costs. It could also be caused by data entry errors. To deal with this data discrepancy, we first aggregated the data from all countries into 14 regions. Aggregating the data into 14 regions reduced the discrepancy because many discrepancies, for example, within a region canceled out with each other after aggregation. At the same time, we only considered trade between regions and ignored trade within a region. We chose to use the year 2011 data for our analysis, which are the most recent data available with relatively small data discrepancy compared to other years. In the end, we only used export data because the data discrepancy may have resulted from importers reporting less to reduce import duties. Table 4.1 displayed the list of the countries aggregated into each region, the total exports and total imports of wood chip for each region in 2011.

Table 4.1 Total exports and imports of wood chip for each region in 2011. Trade quantity was measured in cubic meters. The column 'Countries' contains the countries that were aggregated into its corresponding region. The total exports and imports for a region is the sum of the exports and imports from all the countries in that region.

Region	**Countries**	**Total Exports** (m^3)	**Total Imports** (m^3)
Central America	Bahamas, Barbados, Belize, Costa Rica, Dominican Republic, Guatemala, Haiti, Honduras, Jamaica, Mexico, Nicaragua, Panama	150	4,126
Canada	Canada	1,033,724	2,004,236
East Asia	China, Democratic People's Republic of Korea, Japan, Republic of Korea	1,173	35,240,325
European Union	Austria, Belgium, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, Netherlands, Poland, Portugal, Romania, Slovakia, Slovenia, Spain, Sweden, United Kingdom	1,006,527	6,625,972

Former Soviet Union	Armenia, Azerbaijan, Belarus, Georgia, Kazakhstan, Kyrgyzstan, Republic of Moldova, Russian Federation, Uzbekistan	2,899,220	5,079
Latin America	Argentina, Bolivia, Brazil, Chile, Colombia, Ecuador, Guyana, Paraguay, Peru, Suriname, Trinidad and Tobago, Uruguay, Venezuela	10,058,072	106,356
Middle East	Bahrain, Iran (Islamic Republic of), Israel, Jordan, Kuwait, Lebanon, Oman, Qatar, Saudi Arabia, Turkey, United Arab Emirates	143	3,190,897
North Africa	Egypt, Libya, Mauritania, Morocco, Niger, Tunisia	21,974	20,871
Oceania	Australia, Fiji, French Polynesia, New Caledonia, New Zealand, Papua New Guinea, Solomon Islands, Tonga, Vanuatu	9,755,830	26,050
Other Europe	Albania, Bosnia and Herzegovina, Iceland, Norway, Switzerland, Ukraine	502,738	1,180,499
South Asia	India, Nepal, Sri Lanka	15,904	946

Southeast Asia	Cambodia, Indonesia, Philippines, Singapore, Thailand, Viet Nam	15,103,185	18,294
Sub-Saharan Africa	Burkina Faso, Cameroon, Congo, Côte d'Ivoire, Democratic Republic of the Congo, Ethiopia, Gambia, Ghana, Guinea, Liberia, Madagascar, Malawi, Mauritius, Mozambique, Nigeria, Rwanda, South Africa, Togo, Uganda, United Republic of Tanzania, Zambia, Zimbabwe	2,501,317	6,106
USA	USA	5,659,235	129,435

A caveat of our data input is that we didn't explicitly consider the bioenergy potential for each region, for example, the forest growth in the US. Instead, our data input is the export and import of wood chip for each region. The reason is that we are not trying to predict whether a region can satisfy the EU's demand for wood chip as bioenergy. Instead, we are trying to predict how the global trade of wood chip would change under various scenarios. Knowing the bioenergy potentials alone would not inform those changes of trade. For example, we believe that our results will not be affected even considering the forest growth in the US due to EU's phytosanitary measures. The EU's requirements for phytosanitary measures have significantly limited the trade of softwood chip from the US to the EU. In fact, our model took this effect into account through model calibration using the year 2011's global wood chip trade data. In the year 2011, the US

exported 64 thousand cubic meters wood chip to the EU, which is only 1% of the US's total export of wood chip in that year. Thus, wood chip from the US's forest growth is unlikely to satisfy EU's demand of bioenergy given EU's phytosanitary measure unless the US can eradicate nematodes in its wood chips in the future. Nonetheless, as we will later discuss in the discussion section, estimating the sustainable bioenergy potential will help inform us whether deforestation will happen or not.

The difference between export values and import values should include the transportation cost, but due to some issues with valuing freight transportation costs such as time lag and customs tax avoidance [84], we needed to calibrate the transportation cost within the model to achieve results that matched reality. Calibration was a natural solution to this problem as transportation cost plays a significant part in the trade of wood chip [84].

4.2.2 Mathematical Model

4.2.2.1 General Model Framework

As mentioned previously, EPA is considering whether to approve pulpwood from whole trees to be used as bioenergy that is eligible for RFS support. Meanwhile, the EU's Renewable Energy Directive is increasing the demand for wood bioenergy for heat and power generation in the EU. We want to study how these energy policies would affect the global trade of wood chip. Given this

application problem and based on the literature review we conducted, we chose to use the spatial price equilibrium model (SPE) and the FAOSTAT data previously described.

We adapted the static SPE to model the global wood chip trade flow for one year [91][92]. Since Samuelson presented the equivalence between SPE and linear programming theory, SPE has been used for modeling regional and international trade in food and forest sectors [93]. Lauri et al. applied a partial equilibrium model which is based on an SPE model to estimate the biomass energy potential in 2050 [94]. The general optimization form of the SPE model is the following:

$$\underset{x_{ij}}{Max}\left\{\sum_{i}\int_{0}^{D_i}\theta_i(D_i)dD-\sum_{j}\int_{0}^{S_j}P_j\left(S_j\right)dS-\sum_{ij}c_{ij}x_{ij}\right\}$$

$$subject\ to$$

$$\sum_{i}x_{ij}\ \geq D_j \tag{4.1}$$

$$\sum_{j}x_{ij}\ \leq S_i$$

$$S_i, D_i, x_{ij} \geq 0,\ \forall i,j$$

Samuelson defined the objective function as "net social payoff." The first term, second term and third term in the objective function respectively represents the consumers' utility, production cost, and transportation cost. The constraints represent the trade flow conservation. Subscript i represents demander i and j represents supplier j. Please see Table 4.2 for description of model variables. When this optimization problem is solved, the outputs are the trade flow x_{ij},

demand prices as the dual variables for the demand constraints, and supply prices as the dual variables for the supply constraints.

Table 4.2 Model variables and parameters. This table contains the list of variables and parameters used in our model and their descriptions.

Symbol	Description
x_{ij}	trade flow from region i to region j
D_j	demand at region j
S_i	supply at region i
$\theta_i(D_i)$	demand function at region i
$P_j(S_j)$	supply function at region j
c_{ij}	unit transportation cost from region i to region j
$E_{s,j,k}$	price elasticity of supply for region k in year j
$S_{j,k}$	supply quantity of wood chip for region k in year j
$P_{j,k}$	supply price of wood chip for region k in year j
$S_{2011,k}$	supply quantity of wood chip for region k in year 2011
$P_{2011,k}$	supply price of wood chip for region k in year 2011
α_{ij}, Γ_{ij}	model calibration parameters between region i and j, derived from first stage linear programming model

4.2.2.2 Model Details

In this section, we describe the details of our underlying model as well as the implications for our adaptation of the general SPE model. While our assumptions are necessary for a global trade analysis given insufficient data, we justify our approach using available evidence and provide information on how the results can change if we had more detailed data.

First, we assumed perfectly inelastic demand for wood chip for our policy analysis. Our study focuses on how supply would be affected by increased demand. So, for our policy scenario analysis, it makes sense to fix demand but not supply and adjust demand for different policy scenarios. While there is no previous study about the demand elasticity of wood chip, Kristöfel et al. found the demand for wood pellets in Austria to be inelastic in the short run using a two-stage least squares regression. Our assumptions are consistent with this result [95].

Second, we assumed a linear supply curve. For a perfectly competitive market, the supply curve is equivalent to the upward-sloping part of marginal cost curve where the marginal cost is larger than the supplier's average variable cost [96]. Therefore, for a linear supply curve, the change in marginal cost is constant as production increases. However, if we use a quadratic supply curve, the change in marginal cost would increase as production increases. The total production cost could be much higher if we used quadratic supply curve at higher levels of production. Normally this is because of limitations of technology and the increasing cost of extracting a resource close to its capacity [97]. As it's unlikely for the production of wood chip to reach its

production capacity in any one region during the year, a linear approximation to the supply curve is justified. More data would have allowed us to determine the functional form better, but a linear supply curve captures the dynamics of supply for one year. We constructed our supply curve using arc supply elasticities, reference supply, and reference supply price based on our data as described in detail in the Model Construction section. There are no comprehensive studies about supply elasticity of wood chip.

Finally, we represented transportation cost as a quadratic function of the traded amount of wood chip. A quadratic transportation cost function means the marginal transportation cost is increasing linearly. Intuitively, this makes sense, as the total amount of wood chip transported increases, the unit price of transporting unit amount of wood chip will increase due to reasons such as the shipping vessels reaching capacity. A quadratic transportation cost also implies there is an optimal amount of wood chip to transport, which occurs when the corresponding excess marginal transportation cost is zero. A marginal transportation cost of zero means it is the most cost-effective to transport that amount of wood chip. In conclusion, a quadratic transportation cost function allows for a good representation of increasing marginal costs while still allowing for analysis and calibration.

Given above assumptions, the first term in the objective function of the general SPE model became constant and can be ignored since we used perfectly inelastic demand, i.e., the demand is constant. Therefore, we are minimizing total production cost plus transportation cost given fixed

demand. Then we only need supply functions $P_i(S_i)$ and transportation costs c_{ij} to construct the model as described in the next sections. Furthermore, we assumed the supply curve is linear and transportation cost is a quadratic function of trade quantity of wood chip. So, the supply function $P_i(S_i)$ is a linear function and a quadratic term for the transportation cost would be added to the objective function in model (4.1) in the model construction section. Table 4.2 contains the list of variables used in our model and their corresponding descriptions.

4.2.2.3 Model Calibration

The goal of calibration is to choose model parameters so that the outputs exactly match or are as close as possible to the observed data. In our case, the transportation cost parameters were perfectly calibrated using 2011 wood chip trade data, to ensure that the factors determining transportation costs are embedded in the calibrated parameters. This includes any contracts, taxes, and other cost components of transporting wood chip. Model calibration is a type of parameter estimation and is also called inverse optimization. Several previous studies solved inverse linear programming problems for purposes of calibration [98][99][100].

We used the positive mathematical programming (PMP) approach for our model calibration [101]. PMP calibrates parameters perfectly, and the calibrated model will produce the same solution as observed data. It has been mainly applied to policy analysis and studied in the field of agricultural economics [102]. PMP is a two-stage process. In the first stage, we solved the original optimization problem (linear programming in our case). At the second stage, we

constructed parameters using duals from the first stage problem and observed data. We added another nonlinear term to the objective function in stage 1 using the constructed parameters to form the objective function for the stage 2 problem. The constraints are the same for two stages. The model at stage 2 will give a solution that is the same as the observed data [103].

4.2.2.4 Model Construction

4.2.2.4.1 Overview

We build our final calibrated SPE model in two steps. First, we calibrated the transportation cost of a basic linear programming transportation model. Then, we added production cost to the calibrated model and relaxed the fixed supply constraints. The detailed model construction proceeds as follows:

4.2.2.4.2 Linear programming transportation model

When demand and supply are constant, then SPE model is equivalent to a linear programming transportation model. So first, we solve a linear programming transportation model as following:

$$\begin{aligned} &\underset{x}{Max} - c^T x \\ &subject\ to \\ &Ax \ \leq b\ (duals\ \hat{\lambda}) \\ &x \geq 0 \end{aligned} \tag{4.2}$$

$$where\ x \in \Re^n, A \in \Re^{m \times n}, b \in \Re^m, c \in \Re^n, \hat{\lambda} \in \Re^n$$

Here, $n = 196, m = 28$ for the entire model. The symbols c ($\$/m^3$) are the unit transportation costs of wood chip between regions. Since we do not have perfect transportation cost data, we start with using export prices as c for the linear part of the transportation cost function that we will calibrate. The parameter b represents the wood chip supply and demand for all the regions. Denote $\hat{\lambda}$ as the dual variables for the constraints at the optimal point.

Define parameter $\hat{\gamma} \in \Re^n$ as:

$$\hat{\gamma} = -c - A^T \hat{\lambda} \tag{4.3}$$

The goal is to find a constant vector $\alpha \in \Re^n$ such that:

$$\hat{\gamma} - \alpha > 0 \tag{4.4}$$

Define parameter $\Gamma \in \Re^{n \times n}$ as a positive definite diagonal matrix:

$$\Gamma = diag\left[\frac{(\hat{\gamma}_1 - \alpha_1)}{\hat{x}_1}, \dots, \frac{(\hat{\gamma}_i - \alpha_i)}{\hat{x}_i}, \dots, \frac{(\hat{\gamma}_{196} - \alpha_{196})}{\hat{x}_{196}}\right] \tag{4.5}$$

where $\hat{x}$ is the observed wood chip trade flow.

Build a new quadratic programming model using parameters Γ, α.

$$\begin{aligned} &\underset{x}{Max} - c^T x - \frac{1}{2} x^T \Gamma x - \alpha^T x \\ &subject\ to \\ &Ax \leq b\ (duals\ \lambda) \end{aligned} \tag{4.6}$$

$$x \geq 0$$

$$where\ x \in \Re^n, A \in \Re^{m \times n}, b \in \Re^m, c \in \Re^n$$

When Problem ((4.6) is solved, $x = \hat{x}$ and $\lambda = \hat{\lambda}$. This can be proven by looking at the problem's optimality conditions.

4.2.2.4.3 Add supply function and relax fixed supply

Add supply functions to the objective function and relax fixed supply in the constraint. We first estimated arc elasticities of supply for each region by using the median value of historical arc elasticities. We took supply quantity and supply price for the year 2011 as a reference. For each region, we compute the arc elasticities for each year since 1996 as the following:

$$E_{s,j,k} = \frac{\left(\frac{S_{j,k} - S_{2011,k}}{S_{j,k} + S_{2011,k}}\right)}{\left(\frac{P_{j,k} - P_{2011,k}}{P_{j,k} + P_{2011,k}}\right)} \tag{4.7}$$

Where $S_{j,k}$ and $P_{j,k}$ are the wood chip supply quantity and supply price in year j for region k. We used US's historical consumer price index data to correct the export prices for inflation. All the prices were adjusted to base on US dollars in year 2011.

Before calculating the elasticities, we filtered out supply prices which are either below \$10/tonne or above \$200/tonne. We consider supply prices out of this range (\$10/tonne to \$200/tonne) to be outliers. Regions that have supply price outliers comprise less than 20% of total volume of wood chip exports worldwide from the year 1997 to 2011 in our dataset. These price

outliers are very likely to be incorrect, so to obtain a robust supply elasticity estimate, we excluded these price outliers.

We use the average value of $E_{s,j,k}$ as the supply elasticity for that region k which we denote as $E_{s,k}$.For regions that have negative elasticities, we set their elasticities to be the smallest positive supply elasticities among other regions. The estimated supply elasticities are in Table 4.3.

We constructed a linear supply function for region k using estimated elasticity and reference (the year 2011) supply quantity and price:

$$S_k = S_{2011,k}[1 + E_{s,k}\left(\frac{P_k}{P_{2011,k}} - 1\right)] \tag{4.8}$$

The inverse supply function is:

$$P_k = P_{2011,k}[1 + \frac{1}{E_{s,k}}\left(\frac{S_k}{S_{2011,k}} - 1\right)] \tag{4.9}$$

Note $P_{2011,k}$ is the supply price for region k from the above quadratic programming model i.e., the dual corresponding to the supply constraint in the quadratic programming model. By constructing such supply functions, we can relax fixed supply constraint and add production cost to the objective function without changing the solutions from the calibrated quadratic programming model.

Another simple approach to construct a linear supply curve is to fit a linear regression curve to the supply quantity and price data of wood chip. However, due to an insufficient amount of data (we only have annual data) and high variability of supply price data, such an approach was not

possible. The availability of more wood chip trade data, for example, monthly data, would help produce a more robust linear supply curve using this approach.

Table 4.3 Estimated wood chip supply elasticities for each region

Region	**Supply Elasticities**
Central America	0.27
Canada	1.85
East Asia	0.27
European Union	2.76
Former Soviet Union	5.56
Latin America	4.58
Middle East	0.27
North Africa	0.27
Oceania	0.39
Other Europe	1.10
South Asia	0.27
Southeast Asia	5.30
Sub-Saharan Africa	0.57
USA	0.27

4.2.2.4.4 Final SPE model

Finally, our calibrated adapted SPE model is the following:

$$\underset{x_{ij},\, S_j}{Max} -\sum_j \int_0^{S_j} P_j(S_j)\, dS - \sum_{ij} c_{ij}x_{ij} - \sum_{ij} \frac{1}{2}\Gamma_{ij}x_{ij}^2 - \sum_{ij} \alpha_{ij}x_{ij}$$

$$subject\ to$$

$$\sum_i x_{ij} = D_j \tag{4.10}$$

$$\sum_j x_{ij} \leq S_i$$

$$S_i, x_{ij} \geq 0,\ i,j \in \{1,2,\dots,14\}$$

4.3 Scenarios

4.3.1 Overview

The objective function is minimizing the production cost plus transportation cost, the same as maximizing the negative value of the production cost plus transportation cost. Subscript *i* represents region *i* and *j* represents region *j*. The equality constraint means the total imports into region *j* equals to region *j*'s demand. The inequality constraint means the total exports from region *j* can't exceed its supply. The detailed description of each model variable is in Table 4.2.

We considered two types of scenarios: first, the U.S. decreases its supply to other countries to satisfy its increasing domestic demand for cellulosic biofuel and biomass electricity, and second, the EU increases its demand for renewable energy. To model the first type of scenario, we set an

upper bound for the U.S.'s supply. We obtained the upper bound by reducing the U.S.'s supply for the base case by the projected amount, which is the U.S.'s increased demand for that scenario. For the second type of scenario, we simply increased the EU's demand to the projected level. For each type of scenario, there are different levels of demand and supply given by our projections for the year 2022.

Our projections are based on expected outcomes from two major policies: the EU's Renewable Energy Directive (RED) and the U.S. Renewable Fuel Standard. The first policy incentivizes biomass heat and power as a compliance option for EU's renewable energy mandate. The U.S. Renewable Fuel Standard could incentivize pulpwood biofuel as an eligible compliance option if this pathway is approved. We also consider the growing demand for biomass in electricity production in the U.S. For the EU RED, the total expected heat and power demand is taken from National Renewable Energy Action Plans for all EU Member States. Our scenarios assume that 40% of the target for renewable heat and power in EU is met with wood chip, and subtract estimated consumption of biomass in 2014 to project the demand increase from the present. The EU RED is binding through the year 2020, and we assume constant levels from 2020 to 2022 to be able to compare with the U.S. RFS outcomes. For the U.S. RFS, we assume total cellulosic biofuel production of 1 billion gallons in 2022, which is roughly consistent with the current growth of the industry from 2013-2016 according to historical production and EPA's projections, and that 25% of total cellulosic biofuel is produced from wood chip. We assume a cellulosic ethanol yield

of 105.7 gallons per tonne biomass, based on a futuristic yield from Bloomberg New Energy Finance [103].

Note that our current scenario analysis is based on possible implementation pathways of existing policies (e.g., RFS), not bioenergy in general. The possible pathway is using wood chip as a bioenergy feedstock. Our scenarios analysis can be cited as strategies to help meet broader energy and climate policies such as Paris Agreement goals, which doesn't specifically address how bioenergy should be used to meet the target.

After running our model under different scenarios, we looked at the resulting supply from each region and compared it to the base case. We were specifically interested in which regions increase their supply significantly in each scenario.

4.3.2 Specific Scenarios

In Figure 4.1 and Figure 4.2, we show the results for the following five specific scenarios:

1. The U.S. increases its demand for wood chip by 2.37 million tonnes for biofuel, corresponding to 250 million gallons of pulpwood ethanol in 2022.

2. EU increases its demand for wood chip by 34.78 million tonnes, corresponding to 40% of the EU's renewable energy mandate in 2020-2022.

3. U.S. increases its demand by 2.37 million tonnes for biofuel and EU increases its demand by 34.78 million tonnes, corresponding to 250 million gallons of pulpwood ethanol and 40% of the EU's renewable energy mandate in 2020-2022.

4. U.S. increases its demand by 2.37 million tonnes for biofuel plus 25 million tonnes for power, corresponding to 250 million gallons of pulpwood ethanol and expected growth in U.S. biomass power in 2022.

5. U.S. increases its demand by 2.67 million tonnes for biofuel plus 25 million tonnes for power and EU increases its demand by 34.78 million tonnes, corresponding to 250 million gallons of pulpwood ethanol, expected growth in U.S. biomass power and 40% of the EU's renewable energy mandate in 2020-2022.

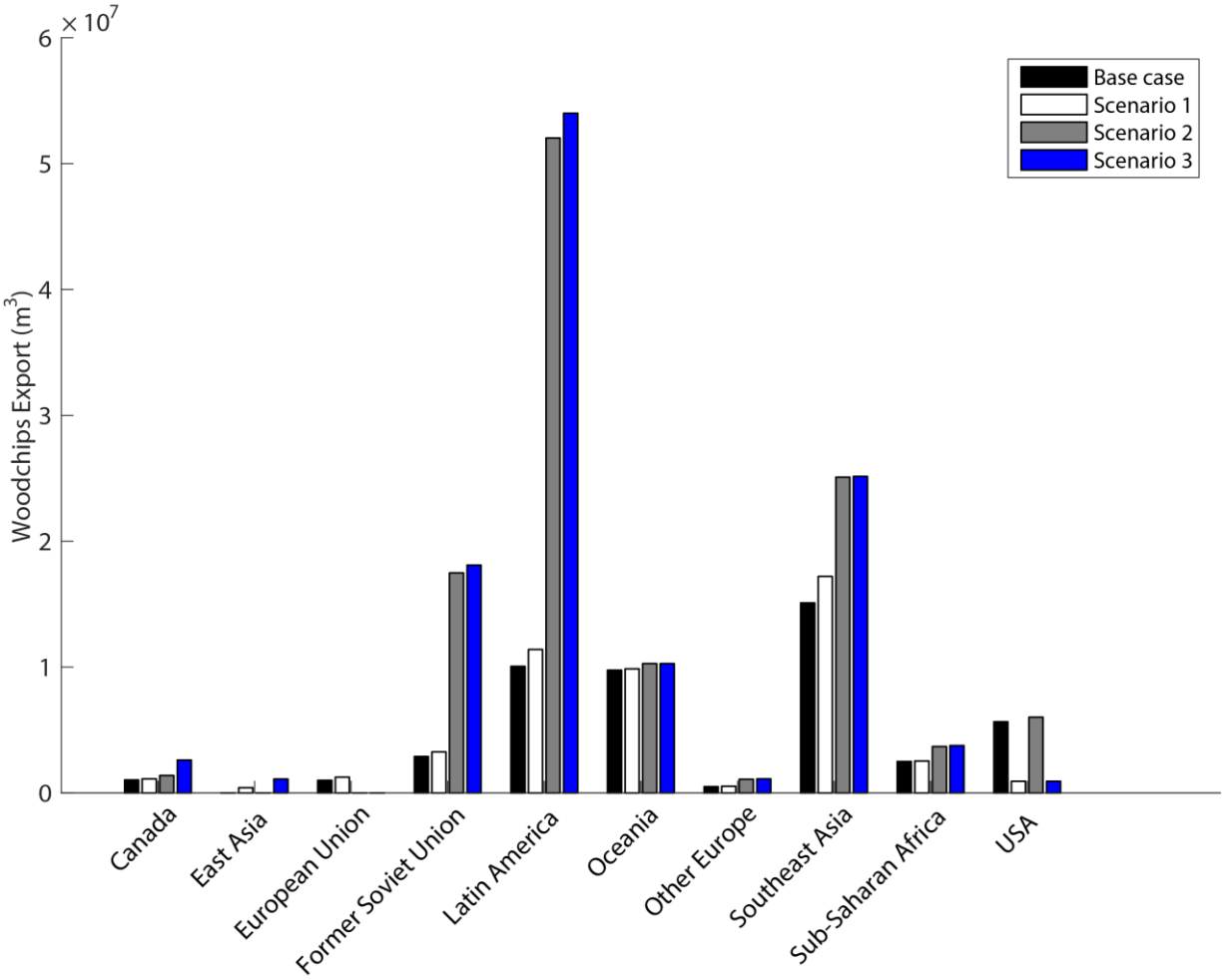


Figure 4.1 Comparison of export of wood chip from different regions between the base case and the first three scenarios. Exports from the Middle East, North Africa, Central America and South Asia were omitted here because they are negligible. Base case: actual exports in the year 2011. Scenario 1: Increase in U.S. demand for cellulosic biofuel. Scenario 2: Increase in EU demand for renewable energy mandate. Scenario 3: Combined demand increase in U.S. and EU.

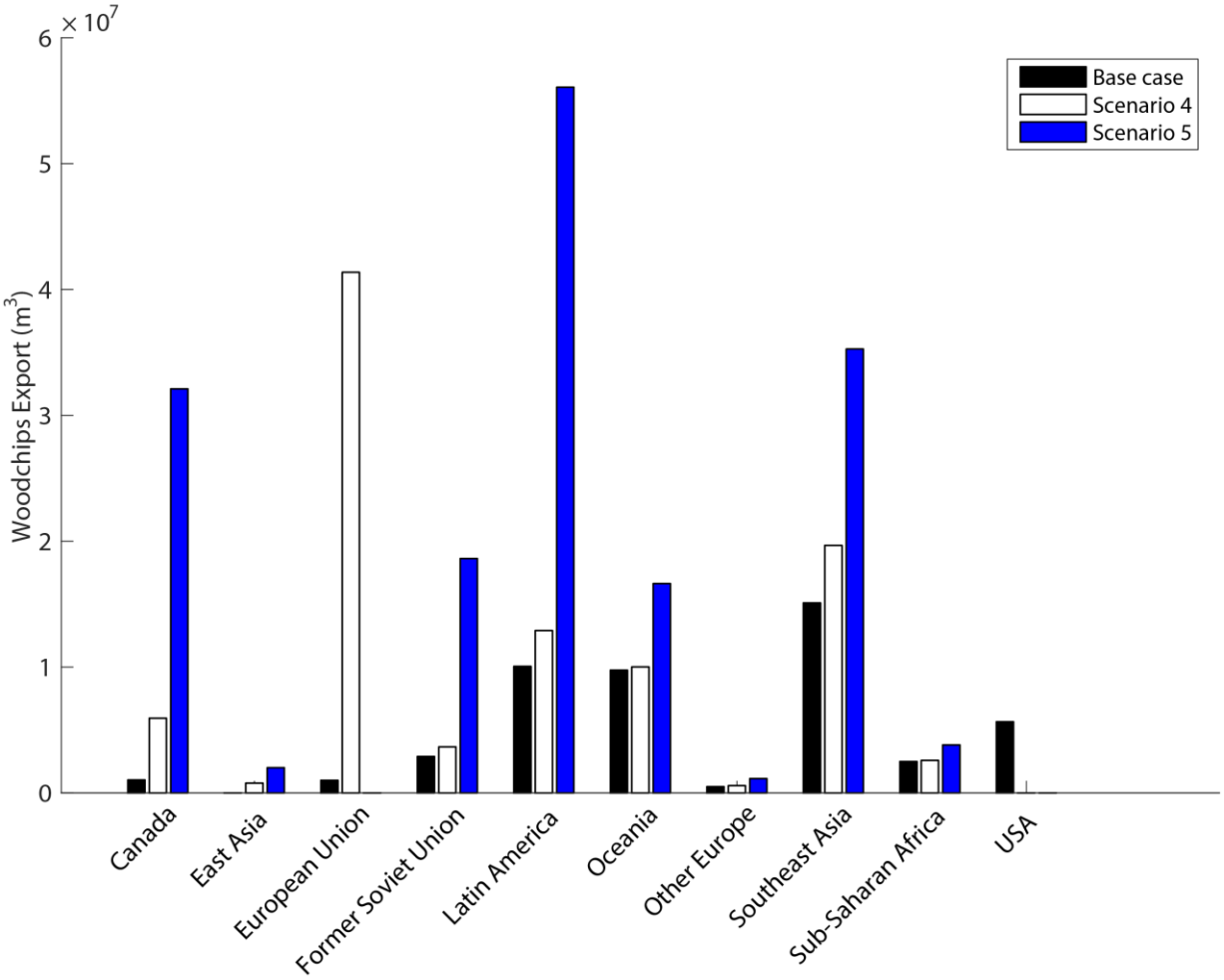


Figure 4.2 Comparison of export of wood chip from different regions between the base case and the fourth and fifth scenarios. Exports from the Middle East, North Africa, Central America and South Asia were omitted here because they are negligible. Base case: actual exports in 2011. Scenario 4: Increase in U.S. demand for cellulosic biofuel and biomass power. Scenario 5: Combined increase in U.S. demand for cellulosic biofuel and biomass power and EU demand for renewable energy mandate.

4.4 Results

4.4.1 Base Case Results

The base case represents the actual trade of wood chip between regions in the year 2011. Table 4.4 displays the trade between major exporters and major importers and a summary of the major export and import data for that year. The largest flow of wood chip between regions is from Southeast Asia to East Asia. Southeast Asia, mainly Vietnam and Thailand had boosted their export since 2010 to satisfy East Asia, especially China's fast-growing demand for wood chip for paper and pulp production. Oceania is the second largest wood chip supplier for East Asia and has been a major exporter worldwide since 1997 because of the demand from Australia. The EU mainly imports wood chip from Latin America and the Former Soviet Union. The Middle East also has been a major importer, largely due to a recent increase in Turkey's wood chip demand. Besides the year 2011, our analysis using the year 1997 to 2011 data shows that historically, Latin America, North America, South East Asia and Oceania have been major exporters of wood chip and the major importers are South East Asia and European Union. Hillring also identified those regions strong in international trade of wood fuel using the year 2000 to 2002 data [88].

Table 4.4 Major trade of wood chip in the year 2011. Trade quantity was measured in thousand cubic meters. Row names represent major exporters and column names represent major importers. The column and row 'Percentage' represent each region's percentage of global imports or exports.

Major Regions	**East Asia** ($K\ m^3$)	**EU** ($K\ m^3$)	**Middle East** ($K\ m^3$)	**Canada** ($K\ m^3$)	**Other Europe** ($K\ m^3$)	**Total Export** ($K\ m^3$)	**Percentage (%)**
Southeast Asia	15,103	0.277	0.075	0.001	0.001	15,103	31.1
Latin America	5,820	3,479	424	0.001	314	10,058	20.7
Oceania	9,736	0.693	0.001	0.001	1000	9,757	20.1
USA	1,517	64	1,989	2,004	0.194	5,659	11.7
Former Soviet Union	207	2,691	1	0.001	0.246	2,899	6.0
Sub-Saharan Africa	2,258	189	0.001	0.001	0.001	2,501	5.2
Total Imports	35,240	6,626	3,191	2,004	1,180		
Percentage (%)	72.6	13.6	6.6	4.1	2.4		

One obvious characteristic of these trade flows is that exporters tend to supply wood chip to geographically closer regions. Even though Southeast Asia is a major exporter and the EU is a major importer, there is almost no trade from Southeast Asia to the EU. The reason is very likely that the shipping cost from Southeast Asia is too high. Intuitively, these trade flows make sense because trade between closer regions has smaller transportation cost.

Another characteristic is that the U.S. and Latin America's exports are more dispersed to different regions. Both regions export significant quantities to East Asia and the EU, and also to the Middle East, and Latin America exports large quantities of wood chip to other European countries.

4.4.2 Scenario Results

Figure 4.1 and Figure 4.2 shows the export of wood chip for scenario 1 to 5 compared with the base case for each region. Figure 4.3 and Figure 4.4 shows the major changes of trade flow of wood chip for scenario 1 and scenario 3. The quantitative scenario results for major changes of trade flow of wood chip are shown in Table 4.5 and Table 4.6.

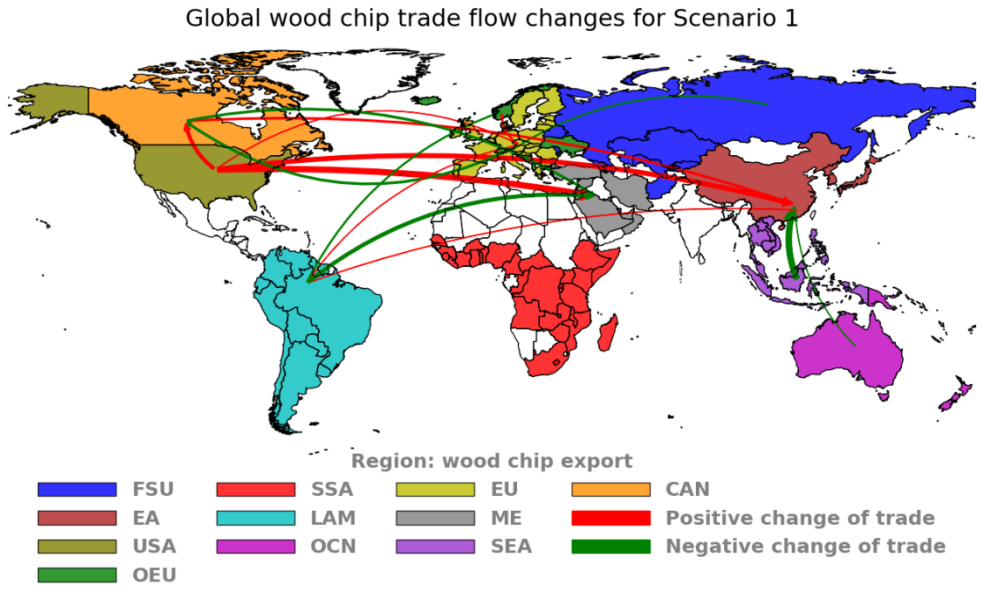


Figure 4.3 Global wood chip trade flow changes for scenario 1. Red arrows show an increase and green arrows show a decrease in the trade as results for scenario 1 when compared to the base case. The width of the arrow represents the relative magnitude of the trade flow changes. Please see Table 5 for the actual values. Here 11 regions were filled with different colors. The other three regions have negligible trade changes and were not color-filled.

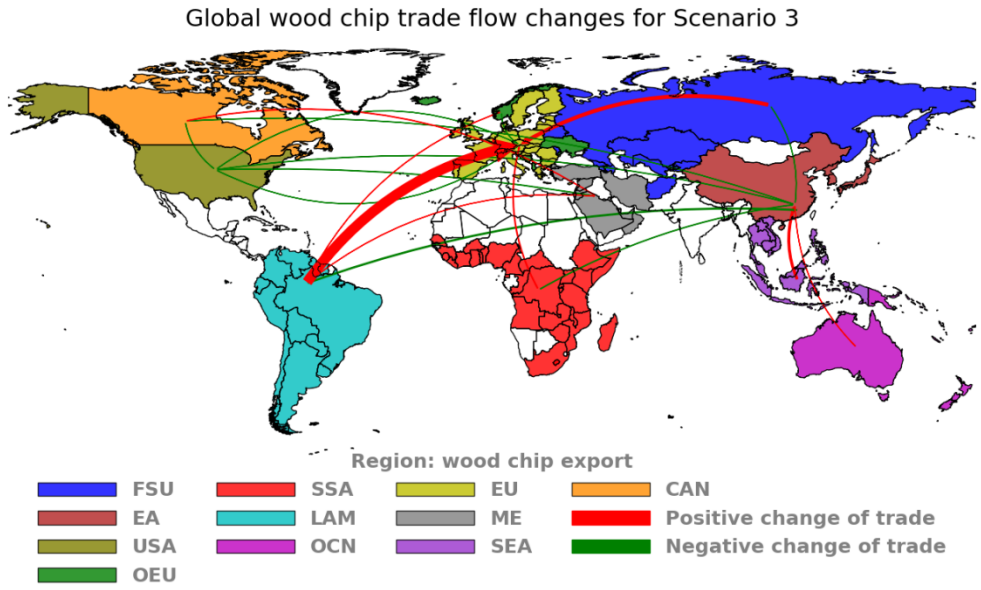


Figure 4.4 Global wood chip trade flow changes for scenario 3. Red arrows show an increase and green arrows show a decrease in the trade as results for scenario 3 when compared to the base case. The width of the arrow represents the relative magnitude of the trade flow changes. Please see Table 6 for the actual values. Here 11 regions were filled with different colors. The other three regions have negligible trade changes and were not color-filled.

Table 4.5 Major changes in trade between major regions for the first scenario compared to the base case in the year 2011. Trade quantity was measured in thousand cubic meters. Row names represent exporters and column names represent importers. Positive numbers mean trade increased compared to the base case and vice versa.

Major Regions	**East Asia** ($K\ m^3$)	**European Union** ($K\ m^3$)	**Middle East** ($K\ m^3$)	**Canada** ($K\ m^3$)	**Other Europe** ($K\ m^3$)	**USA** ($K\ m^3$)
Southeast Asia	2,096	0.019	0.031	0.006	0	0.008
European Union	-1	0	-4	655	-406	-26
Latin America	-132	-80	1,141	0.03	404	0
Oceania	96	0	0	0.006	3	7
USA	-1,517	-64	-1,989	-1,077	-0.194	0
Former Soviet Union	37	324	1	0.006	0.2	0
Sub-Saharan Africa	-9	-0.743	0	0.006	0	0
Canada	-570	-0.001	644	0.006	0	11

Table 4.6 Major changes in trade between major regions for the third scenario compared to the base case in the year 2011. Trade quantity was measured in thousand cubic meters. Row names represent exporters and column names represent importers. Positive numbers mean trade increased compared to the base case and vice versa.

Major Regions	**East Asia** ($K\ m^3$)	**European Union** ($K\ m^3$)	**Middle East** ($K\ m^3$)	**Canada** ($K\ m^3$)	**Other Europe** ($K\ m^3$)	**USA** ($K\ m^3$)
Southeast Asia	10,037	8	0.475	0.015	0	0.108
European Union	-2	0	-25	-0.102	-865	-105
Latin America	-5,820	48,758	183	0.077	818	0
Oceania	365	2	0	0.015	47	105
USA	-1,517	-64	-1,989	-1,102	-0.194	0
Former Soviet Union	-207	15,412	-1	0.014	-0.034	0
Sub-Saharan Africa	-2,258	3,469	0	0.015	0	0
Canada	-598	3	2,161	0.015	0	8

From Figure 4.1, we can see that for the first three scenarios, the main regions that provide the extra supply of wood chip are Latin America, Former Soviet Union, and Southeast Asia.

Table 4.5 displays the trade flow changes when only the U.S. increases its demand for wood chip by 2.4 million tonnes for biofuel. Increasing demand will limit the U.S.'s export to other countries, as it will only consume its domestic supply. As a result, other major wood chip importers will have to increase their imports from regions other than the U.S. In this scenario, the Middle East will increase its import from Latin America instead of the U.S. to satisfy its demand. Southeast Asia will increase its export to East Asia in place of the U.S. The Former Soviet Union will increase its export to EU in place of the U.S.

Table 4.6 shows that if the EU and the U.S. both increase their domestic demand, exports from Latin America and the Former Soviet Union to the EU will soar. At the same time, Southeast Asia's exports to East Asia will also increase significantly because all the other major exporters except Oceania will shift their export from East Asia to the EU.

Figure 4.2 shows that if the U.S. further increases its demand by 25 million tonnes for power, in addition to the 2.4 million tonnes domestic demand increase for biofuel, the U.S. won't be able to satisfy its demand and will become a net importer of wood chip. In this scenario, Canada's export to the U.S. will increase significantly. Figure 4.2 also shows greatly increased exports from Latin America and Southeast Asia, and to a lesser extent, from the Former Soviet Union, to satisfy the combined mandates in the U.S. and EU (Scenario 5).

Increased export from Latin America and Southeast Asia may have a negative environmental impact in these regions, especially those countries with differing forest protection

policies and enforcement. For instance, Chile has already been experiencing a clearing of natural forest and plantation expansion for the past few decades [104], partly caused by increasing demand for timber and fuel wood product. Our data analysis also shows Chile has constantly been a major exporter of wood chip from 2001 to 2010, its export of wood chip is about 9% of global export of wood chip annually. So increasing exports of wood chip from Latin America may cause further clearing of natural forest and plantation expansion in Chile, which would likely have negative impacts on the environment and biodiversity [105]. Another example is Sumatra, Indonesia, where increased value of agricultural commodities eroded enforcement efforts and led to increased deforestation in the late 1990s [106], illustrating that increasing demand for forest products can put pressure on forest protection efforts. Also, there is already evidence showing that UK's demand for wood pellets has threatened the wetland forests in the southern U.S. [107]. A Freedom of information request by environmental organization Biofuelwatch showed that British utility company Drax power requires wood from slow-growing trees, not forest residues or energy crops as its source of wood pellets [108][109].

4.5 Discussion

Our results show that if a significant fraction of the cellulosic biofuel mandate under the RFS is met by pulpwood biofuel, this pathway would likely have indirect effects on the global wood market, leading to increased wood harvesting in vulnerable tropical nations. Our model isolates

these demand shocks and does not account for potential demand reduction for wood chip, and so we may overestimate the market effects of the RFS and RED; however, our analysis is illustrative of the type of effects that are likely to occur.

Our results indicate that, to ensure sustainable development of wood chip bioenergy, environmental, social and economic sustainability criteria should be implemented especially in these potential major sourcing regions for bioenergy. On the supply side, in 2011, Janssen and Rutz identified that no specific biofuel sustainability certification system had been implemented, but several sustainability initiatives have been established by stakeholders and governmental bodies from Latin America [110]. On the demand side, EU-wide sustainability schemes and criteria exist for biofuel but not for biomass [71][111]. If sustainability criteria and a certification system regarding wood chip are implemented in these sourcing regions and if we can estimate the sustainable wood chip bioenergy potential, we may be able to tell whether deforestation will occur given the predicted scenario results from our model. More specifically, if the sustainable wood chip bioenergy potential from these regions can't satisfy the increased demand for those scenarios and the sustainability scheme is not implemented in these regions, then theoretically it could lead to deforestation. If the sustainable wood chip bioenergy potential can't satisfy the increased demand and sustainability criteria are implemented in these regions, deforestation may still happen due to potential leakage. For example, let's say sustainable wood can be used for both furniture manufacturing and bioenergy. Due to a higher price for sustainably certified wood for bioenergy

and lack of explicit sustainability criteria for furniture, the sustainable plantation owner will sell wood for bioenergy while cutting down non-sustainably certified wood for furniture. Therefore, implementation of sustainability criteria for bioenergy can reduce, but not eliminate unsustainable deforestation. Therefore, to predict whether increased demand for wood chip from the EU and US will cause deforestation in these major sourcing regions, we need to consider the sustainability constraints regarding wood chip, the amount of sustainable wood chip bioenergy from the sourcing regions and the potential leakage effect. Lamers et al. studied the impact of sustainability criteria on potential import and supply costs of global solid biomass trade to North-West Europe [112]. Their approach incorporated sustainability criteria using feedstock exclusion. However, this approach is not applicable to our analysis since we are studying a single type of feedstock. Incorporating sustainability constraints into our model will also lead to more realistic results in the future when the sustainability criteria are implemented in these major sourcing regions. However, estimating sustainable wood chip bioenergy potential is beyond the scope of this study. Nonetheless, our model can use the results from other studies to estimate sustainable wood chip bioenergy potential as input and answer the question regarding deforestation.

Our results are dependent on the value of supply elasticities, but our methodology provides an approach to make an inference from limited and noisy wood chip trade data. Better data and better estimates of supply elasticities will allow for deeper insights. For example, Kristöfel et al. applied a two-stage least regression method to estimate the demand and supply elasticities of wood

pellet in Austria by constructing demand and supply models [96]. Their demand and supply models include factors that affect both the quantity and price of wood pellet demand and supply such as the number of total installed pellet boilers, heating degree days and production capacity. These methods can be used to estimate elasticities with relevant data for wood chips as well.

Meanwhile, technological, economic and policy factors regarding wood chip bioenergy have been changing such as combustion technology, quality standards, shipping costs, oil prices and phytosanitary rules. In future studies, we can include these factors as constraints and parameters into our model. For example, if we know quantitatively how much combustion technology has improved the efficiency of using wood chip as bioenergy, we can adjust the demand for wood chip accordingly in our model.

For future work, we propose a more nuanced representation of supply and demand. Detailed representation of the availability of wood chip through analysis of forest area, governance quality and production profiles will grant further validation to our results. If data of wood chip directly traded as bioenergy is available in the future, we can use that data and better estimate the effect of bioenergy policy on global trade of wood chip. Finally, a multi-period analysis will give better indications of trends over time. Our model is also amenable to be coupled with other policy models for the renewable fuel standard [113][114] [115], and can thus allow for a more robust policy analysis about climate change mitigation [116][117][118].

4.6 Conclusion

Countries in Latin America, Southeast Asia, and the Former Soviet Union have great amounts of forest resource compared to other countries. Our study quantitatively shows that increased demand for wood chip from the U.S. and EU driven by a combination of renewable energy policies would increase harvests in these countries. Our methodology helps us answer the counterfactual questions for different bioenergy policy scenarios given limited data. This will assist policymakers to make sustainable bioenergy policies. If these countries have poor management and regulation of their forest resources, this may lead to the unsustainable development of wood chip bioenergy with negative impacts on carbon stocks, biodiversity, and rights of indigenous people. As a result, increased demand for wood chip from new renewable energy policies, including a pulpwood biofuel pathway under the RFS, may not deliver the full environmental benefits intended by those policies. To study whether these renewable energy policies will lead to the unsustainable development of wood chip bioenergy, further analysis by interviews with industry experts and field studies can be carried out and combined with our model results.

Chapter 5

Hyper-parameter Optimization for Support Vector Machines

5.1 Introduction

Machine learning models usually contain hyper-parameters that need to be optimized to prevent overfitting and minimize generalization error. Currently, the standard approach for assessing the generalization error of machine learning models is cross-validation [2][119]. The standard practice to optimize the hyper-parameters is using grid-search or random search combined with cross-validation. Specifically, these methods search through a fixed (grid-search) or random (random search) set of values for the hyper-parameters and choose the one that yields the best model performance evaluated by cross-validation. Grid search and random search are computationally

expensive especially given high-dimensional hyper-parameter space. Moreover, a locally optimal solution is not even guaranteed.

Support vector machines (SVM) are popular and efficient methods for supervised learning, specifically, classification tasks [3]. They contain a regularization hyper-parameter C to control for model complexity and overfitting. SVM can also have additional kernel hyper-parameters if we decide to use non-linear kernels, such as a commonly used kernel, Gaussian radial basis function kernels or RBF kernel. In practice, often grid-search or random search is used to choose the hyper-parameters. For more complex machine learning models, particularly, deep neural networks (deep learning models), the training process often involves tuning the hyper-parameter (most importantly, the learning rate, i.e., the step size in gradient descent method) manually while babysitting the training process. As the dimensionality of dataset becomes very large, it's computationally expensive and inefficient to use those empirical ad hoc methods for hyper-parameter tuning. On the other hand, choosing good hyper-parameters is very important to obtain trained machine learning models with good predictive performance on unseen test data. It is often the critical step for building complex nonlinear machine learning models such as deep neural networks, and SVM with nonlinear kernels.

Researchers have started to create more efficient hyper-parameter tuning methods for machine learning algorithms in general. Hyper-parameter optimization has been naturally formulated as a bilevel optimization problem. On the outer or upper level, we are minimizing the

prediction error on unseen test data, also called the generalization error. On the inner or lower level, we are minimizing the error of model fitting on training data. Gradient-based methods for optimizing hyper-parameters have emerged as early as 1999 by Bengio [120]. Those gradient-based methods mostly rely on the implicit differentiation trick. Precisely, those methods differentiate the upper-level loss function with respect to the hyper-parameters using an implicit equation (the optimality condition of inner optimization problem). For instance, Bengio (1999) [120] used a gradient-based method by computing the gradient of the loss function with respect to hyperparameters using the implicit function theorem to optimize multiple hyper-parameters for a general smooth training loss function, particularly the quadratic loss function. Chuong et al. (2007) [121] used gradient-based methods with the implicit function theorem to efficiently optimize multiple hyper-parameters for log-linear models. To avoid computing a demanding exact gradient in the implicit differentiation trick, Pedregosa (2016) [122] optimized the hyper-parameters with approximate gradients and provided numerical results on l_2-regularized logistic regression and kernel Ridge regression models.

Recently, new gradient-based methods have been emerging for hyper-parameter optimization for deep neural network models. These methods make the hyper-parameter tuning process faster and by overcoming expensive memory requirement while retaining good model performance [123][124][125][126]. Typical hyper-parameters for deep neural networks includes learning rates and regularization parameters. Computing the hyper-parameter gradient using

reverse-mode differentiation usually requires iterations of forward and backward pass of computations. It also requires storing the entire training trajectory for millions of intermediate parameters in memory, which is unmanageable. Maclaurin et al. (2015) [123] showed that, instead of storing the entire training trajectory, we could recompute the learning trajectory during backpropagation on the fly by storing few auxiliary bits. They demonstrated the idea for training procedure of stochastic gradient descent (SGD) with momentum. Fu et al. (2016) [124] demonstrated that a shortcut could be created to approximate the reverse-mode differentiation step by extracting the knowledge from the forward pass. Their algorithm is 45 times faster and requires 100 times less memory compared to standard methods evaluated on two image datasets.

SVM has a unique optimization problem structure, which can be explored and utilized to optimize its hyper-parameters. Existing methods mostly optimize the regularization hyper-parameter C, not the kernel hyper-parameter λ due to the nonlinearity and computation complexity induced by λ. Hastie et al. (2004) [127] created a method that can trace the entire regularization path of SVM solutions for different values of the hyper-parameter C. Exploiting the fact that the dual variables of SVM are piecewise-linear in C, their method has the same computational cost as solving one SVM problem. Bennett et al. (2008) [128] and Kunapuli et al. (2008) [129] formulated hyper-parameter optimization as a bilevel optimization problem. They further reduced the problem to a single optimization problem by replacing the lower-level SVM with its optimality conditions, i.e., Karush-Kuhn-Tucker (KKT) conditions [130]. The resulting single optimization problem is

mathematical programming with equilibrium constraints (MPEC). The authors explored various general optimization methods for solving MPEC problems with limited success. The disadvantage of this approach is that the number of constraints grows linearly with the training data size, which makes the optimization problem intractable for large data. Moreover, it doesn't exploit the particular structure of SVM to achieve a more efficient method. Couellan and Wang (2015) [131] took the same bilevel formulation as Bennett et al. (2008) [128] but used SGD with the implicit differentiation trick to optimize the hyper-parameter C. Numerical results showed that their method is efficient for large datasets by exploiting the structure of SVM and performing cheap gradient estimate using SGD. However, SGD itself requires additional hyper-parameters, i.e., the learning rates. As Couellan and Wang's (2015) [131] bilevel-SGD method requires performing SGD both on the upper-level and lower-level problems, it requires two additional hyper-parameters to tune.

To improve the bilevel-SGD method, we further exploited the structure of SVM. Precisely, we adopted a dual coordinate descent method (DCD) [12] for the lower-level problem and combined it with SGD on the upper-level problem. Our approach avoids introducing additional hyper-parameters for the lower-level problem. Numerical results on multiple benchmark datasets show our method converges fast and achieves good generalization performance. First, we present our problem formulation. Then we explain our approach and algorithms used in detail. Finally, we

compare our approach with the bilevel-SGD method and show the empirical results on multiple benchmark datasets.

5.2 Problem Setting

5.2.1 SVM Optimization Problem

For a binary classification problem, let's define the training data as $\{\boldsymbol{x}_i \in R^p,\ i = 1, \dots, N\}$ and the binary labels as $\{y_i \in \{-1,1\},\ i = 1, \dots, N\}$, where p is the number of features and N is the number of samples. SVM is a maximum margin classifier that uses a hyperplane $\{\boldsymbol{x}: f(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\beta} + \beta_0 = 0\}$ to separate the feature space, where $\boldsymbol{\beta}$ and β_0 are the parameters to be learned. The decision function or classification rule is $G(\boldsymbol{x}) = \text{sign}[f(\boldsymbol{x})] = \text{sign}[\boldsymbol{x}^T\boldsymbol{\beta} + \beta_0]$.

For the hard-margin case, where the data is linearly separable, the SVM optimization problem is:

$$\begin{aligned} &\underset{\boldsymbol{\beta},\beta_o}{Min}\ \|\boldsymbol{\beta}\| \\ &subject\ to \quad y_i\ (\boldsymbol{x}_i\boldsymbol{\beta} + \beta_0) \geq 1,\ \ i \in \{1, \dots, N\} \end{aligned} \tag{5.1}$$

where $\frac{\mathbf{1}}{\|\boldsymbol{\beta}\|}$ is the margin size between the data points and the separating hyperplane. The constraints ensure the training data was classified to be on the correct side of the two margins, i.e. $|(\boldsymbol{x}_i\boldsymbol{\beta} + \beta_0)| = 1$. The two margins are two separating hyperplanes $\{\boldsymbol{x}: (\boldsymbol{x}\boldsymbol{\beta} + \beta_0) = 1\}$ and

$\{\boldsymbol{x}: (\boldsymbol{x\beta} + \beta_0) = -1\}$ that separates the training data into two classes and $\frac{\mathbf{2}}{\|\boldsymbol{\beta}\|}$ is the margin size between the two separating hyperplanes.

For the linearly non-separable, or soft margin case, hinge loss is introduced into Eq. (5.1). The optimization problem becomes minimizing the sum of the hinge loss and the inverse of margin size as follows:

$$\underset{\boldsymbol{\beta},\beta_o}{Min}\ \frac{1}{2}\|\boldsymbol{\beta}\|^2 + C\sum_{i=1}^{N}\max\{0, 1 - y_i(\boldsymbol{x}_i\boldsymbol{\beta} + \beta_0)\} \tag{5.2}$$

where $\max\{0, 1 - y_i(\boldsymbol{x}_i\boldsymbol{\beta} + \beta_0)\}$ is the hinge loss for classifying the training example. The loss is positive when the training example is incorrectly classified, i.e., when $y_i(\boldsymbol{x}_i\boldsymbol{\beta} + \beta_0) < 1$. Figure 5.1 demonstrates how a soft margin SVM separates two-dimensional feature spaces into two classes using a hyperplane learned after training.

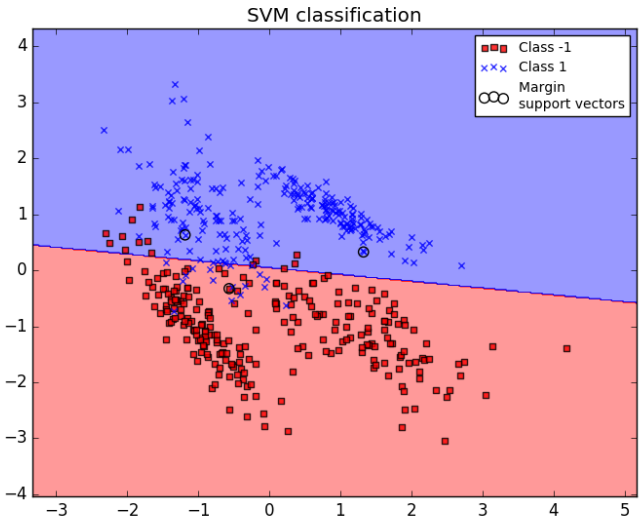


Figure 5.1 A linear support vector machine (SVM) classifier separates a two-dimensional dataset (simulated data) into two classes using a hyperplane learned after training.

Note $\|\boldsymbol{\beta}\|$ in (5.1) was changed to $\frac{1}{2}\|\boldsymbol{\beta}\|^2$ in (5.2) for computational convenience and it doesn't change the optimization problem. The decision function for prediction is still $G(\boldsymbol{x})$. Various SGD methods have been proposed to efficiently solve formulation (5.2) for large-scale datasets [132][133].

The hinge loss function in (5.2) makes the objective function non-smooth. Classically, the problem (5.2) has been reformulated into a constrained quadratic programming problem by introducing a slack variable ξ_i. The optimization problem can be reformulated as:

$$\begin{aligned} & \underset{\boldsymbol{\beta},\beta_o,\xi_i}{Min}\ \frac{1}{2}\|\boldsymbol{\beta}\|^2 + C\sum_{i=1}^{N}\xi_i \\ subject\ to\quad & y_i\,(\boldsymbol{x}_i\boldsymbol{\beta}+\beta_0) \geq 1-\xi_i \quad (dual\ variable: \alpha_i) \\ & \xi_i \geq 0,\ \ i \in \{1,\dots,N\} \quad (dual\ variable: \gamma_i) \end{aligned} \tag{5.3}$$

5.2.2 SVM Dual Problem

Eq. (5.3) is the primal problem for SVM. Its dual problem is a more straightforward quadratic programming problem that only involves one set of decision variables α_i. The dual problem is typically solved instead. Efficient decomposition and coordinate descend methods exits for solving it with large-scale datasets [134][12]. The constraints for the primal problems are all linear constraints, and its objective function is convex. Therefore, Slater's condition is satisfied which implies strong duality holds [135]. As a result, we can solve the dual problem and obtain the same solution as if we solve the primal problem. Moreover, the KKT conditions are sufficient and necessary for optimality. We derive the dual problem from the Lagrangian function as follows:

$$\mathcal{L}(\boldsymbol{\beta},\beta_0,\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{\gamma}) = \frac{1}{2}\|\boldsymbol{\beta}\|^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i[y_i\,(\boldsymbol{x}_i\boldsymbol{\beta}+\beta_0) - 1 + \xi_i] - \sum_{i=1}^{N}\xi_i\gamma_i \tag{5.4}$$

The dual problem is:

$$\begin{aligned} & \underset{\boldsymbol{\alpha},\boldsymbol{\gamma}}{Max}\ \underset{\boldsymbol{\beta},\beta_0,\boldsymbol{\xi}}{min}\ \mathcal{L}(\boldsymbol{\beta},\beta_0,\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{\gamma}) \\ subject\ to\quad & \boldsymbol{\alpha},\boldsymbol{\gamma} \geq \mathbf{0} \end{aligned} \tag{5.5}$$

First, to obtain $\min_{\boldsymbol{\beta},\beta_0,\boldsymbol{\xi}} \mathcal{L}(\boldsymbol{\beta},\beta_0,\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{\gamma})$, we set the derivative of $\mathcal{L}(\boldsymbol{\beta},\beta_0,\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{\gamma})$ with respect to $\boldsymbol{\beta},\beta_0,\boldsymbol{\xi}$ to 0.

$$\frac{\partial \mathcal{L}(\boldsymbol{\beta},\beta_0,\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{\gamma})}{\partial \boldsymbol{\beta}} = \boldsymbol{\beta} - \sum_{i=1}^{N} \alpha_i y_i \boldsymbol{x_i^T} = 0 \tag{5.6}$$

Therefore, we have $\boldsymbol{\beta} = \sum_{i=1}^{N} \alpha_i y_i \boldsymbol{x_i^T}$.

$$\frac{\partial \mathcal{L}(\boldsymbol{\beta},\beta_0,\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{\gamma})}{\partial \beta_0} = -\sum_{i=1}^{N} \alpha_i y_i = 0 \tag{5.7}$$

$$\frac{\partial \mathcal{L}(\boldsymbol{\beta},\beta_0,\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{\gamma})}{\partial \xi_i} = C - \alpha_i - \gamma_i = 0 \tag{5.8}$$

Combing Eq. (5.6), Eq. (5.7), Eq. (5.8) into Eq. (5.4) and simplifying, we obtain:

$$\mathcal{L}(\boldsymbol{\beta},\beta_0,\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{\gamma}) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \boldsymbol{x_i^T x_j} \tag{5.9}$$

Combing Eq. (5.5), (5.7), (5.8), (5.9), we obtain the dual problem for SVM:

$$\begin{aligned} &\max_{\alpha_i} \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \boldsymbol{x_i^T x_j} \\ &subject\ to \quad 0 \le \alpha_i \le C \\ &\sum_{i=1}^{N} \alpha_i y_i = 0\,, \quad i,j \in \{1,\dots,N\} \end{aligned} \tag{5.10}$$

Platt (1998) created one of the first decomposition methods, sequential minimal optimization (SMO), to efficiently solve the dual problem [134]. Later Hsieh et al. (2008) proposed a dual coordinate descent method to solve the slightly reformulated problem in (5.12). Their approach is well suited for large-scale linear SVM problems [12]. It takes $O(\log\left(\frac{1}{\epsilon}\right))$ iterations for their DCD method to reach an ϵ-accurate solution.

By including the bias term β_0 into the vector $\boldsymbol{\beta}$ and append another constant column filled with 1 to the input data, the problem (5.10) can be further simplified [12]. Specifically, after the following operation:

$$\boldsymbol{x}_i^T \leftarrow [\boldsymbol{x}_i^T, 1], \quad \boldsymbol{\beta}^T \leftarrow [\boldsymbol{\beta}, \beta_0] \tag{5.11}$$

The dual problem is reformulated as follows:

$$\begin{aligned} &\underset{\alpha_i}{Max} \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j \\ &= \boldsymbol{e}^T\boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}^T Q \boldsymbol{\alpha} \\ &subject\ to \quad 0 \le \alpha_i \le C, \quad i \in \{1, \dots, N\} \end{aligned} \tag{5.12}$$

where $Q_{ij} = y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j$, and $\boldsymbol{e}$ is a unit vector.

5.3 Bilevel Problem Formulation

In general, a bilevel problem is expressed as follows:

$$\begin{aligned} & \underset{x,\bar{y}}{Min}\ F(x,\bar{y}) \\ subject\ to\quad & G(x,\bar{y}) \le 0 \\ & \bar{y} = \underset{x,\bar{y}}{argmin}\{f(x,y): g(x,y) \le 0\} \end{aligned} \tag{5.13}$$

where $\bar{y}$ is the optimal solution for the lower level optimization problem and x and $\bar{y}$ both decision variables for the upper-level optimization problem.

5.3.1 One-Fold Validation

Usually, the hyperparameter is tuned by maximizing model performance on the hold-out dataset using k-fold cross-validation [119]. Let's first consider a simple case using the same framework where we optimize the hyperparameter to maximize the model prediction performance on one separate validation set (N samples). Precisely, on the upper level, we are minimizing prediction loss on the validation set. On the lower level, we are solving the SVM problem itself. Let's again ignore the bias term β_0 using operation (5.11) and take the original SVM primal problem formulation, Eq. (5.2), as the lower-level problem. The bilevel problem formulation for hyper-parameter optimization for SVM becomes:

$$\begin{aligned} & \underset{C,\bar{\boldsymbol{\beta}}}{Min}\ F\left(C,\bar{\boldsymbol{\beta}}\right) = \sum_{j=1}^{N} \max\{0, 1 - y_j\left(\bar{\boldsymbol{\beta}}^{\boldsymbol{T}}\boldsymbol{x}_j^{v}\right)\} \\ subject\ to\quad & C_{\max} \ge C \ge 0 \\ & \bar{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{argmin}\{G(C,\boldsymbol{\beta}) = \frac{1}{2}\|\boldsymbol{\beta}\|^2 + C\sum_{i=1}^{L} \max\{0, 1 - y_i(\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{x}_i)\}\ \} \end{aligned} \tag{5.14}$$

where $\boldsymbol{x}_j^v$ represents data in the validation set, and $\boldsymbol{x}_i$ represents data in the training set.

5.3.2 Related Work

Couellan and Wang (2015) solved the above problem (5.14) using SGD [131]. The key challenge is to use SGD is to obtain the gradient estimate of $F\left(C, \overline{\boldsymbol{\beta}}\right)$ with respect to hyper-parameter C, $\nabla_C F(C, \overline{\boldsymbol{\beta}}(C))$. To deal with this, they used the implicit differentiation trick and chain rule for deriving derivatives.

Using chain rule, we obtain:

$$\nabla_C F\left(C, \overline{\boldsymbol{\beta}}(C)\right) = \nabla_C F\left(C, \overline{\boldsymbol{\beta}}\right)^T + \nabla_{\overline{\boldsymbol{\beta}}} F\left(C, \overline{\boldsymbol{\beta}}\right)^T \nabla_C \overline{\boldsymbol{\beta}}(C) \tag{5.15}$$

When the lower level problem reaches the optimum solution, the sub-gradient equals zero, i.e.,

$$\nabla_{\boldsymbol{\beta}} G(C, \boldsymbol{\beta}) = \boldsymbol{\beta} - \boldsymbol{C} \sum_{l=1}^{L_e} y_l \boldsymbol{x}_l \ = 0 \tag{5.16}$$

where l is the index of the sample whose training loss is positive, i.e., $\{l = 1, \dots, L_e | y_l \boldsymbol{\beta}^{\boldsymbol{T}} \boldsymbol{x}_l < 1\}$, as the for the other correctly classified samples, the sub-gradient is just zero.

They used SGD to solve the lower-level problem and randomly sampled one data instance to estimate the sub-gradient for each iteration. Eq. (5.16) was further approximated by

$$\nabla_{\boldsymbol{\beta}} G_l(C, \boldsymbol{\beta}) = \boldsymbol{\beta} - \boldsymbol{C} y_l \boldsymbol{x}_l = 0 \tag{5.17}$$

Using this optimality condition for the lower-level problem, they applied implicit function theorem to obtain:

$$\nabla_C \overline{\boldsymbol{\beta}}(C) = -\nabla^2_{\boldsymbol{\beta}} G_l(C, \boldsymbol{\beta})^{-1} \nabla^2_{\boldsymbol{\beta C}} G_l(C, \boldsymbol{\beta}) \tag{5.18}$$

From Eq. (5.17) we can derive $\nabla^2_{\boldsymbol{\beta}} G_l(C, \boldsymbol{\beta}) = 1$ and $\nabla^2_{\boldsymbol{\beta C}} G_l(C, \boldsymbol{\beta}) = y_l \boldsymbol{x}_l$. Eq. (5.18) can be simplified as:

$$\nabla_C \overline{\boldsymbol{\beta}}(C) = y_l \boldsymbol{x}_l \tag{5.19}$$

They also used SGD to optimize the upper-level problem and estimated the sub-gradient $\nabla_{\overline{\boldsymbol{\beta}}} F(C, \overline{\boldsymbol{\beta}})$ using one random data instance as an unbiased noisy gradient estimate. More specifically:

$$\nabla_{\overline{\boldsymbol{\beta}}} F\left(C, \overline{\boldsymbol{\beta}}\right) = -y_p^v \boldsymbol{x}_p^v, \text{ where } p \in \{1, \dots, N_e | y_p^v \boldsymbol{\beta}^{\boldsymbol{T}} \boldsymbol{x}_{\boldsymbol{p}}^{\boldsymbol{v}} < 1\} \tag{5.20}$$

Inserting Eq. (5.20) and $\nabla_C F\left(C, \overline{\boldsymbol{\beta}}\right)^T = 0$ into Eq. (5.15), they obtained:

$$\nabla_C F\left(C, \overline{\boldsymbol{\beta}}(C)\right) = -y_p^v \boldsymbol{x}_p^v y_l \boldsymbol{x}_l \tag{5.21}$$

Then they optimized both the hyper-parameter C on the upper level and model parameters on the lower level iteratively using simple SGD [131].

The disadvantage of their approach is that they introduced two new hyper-parameters, the learning rates α_w^t and α_C^t for the upper-level SGD and lower-level SGD parameter updates. The learning rates themselves are hard to tune as well.

5.3.3 Our Approach

To further improve their method, we propose an approach to combine SGD and DCD methods to optimize the hyper-parameter C. We choose to optimize the SVM dual problem formulation on the lower level using a DCD method proposed by Hsieh et al. (2008), while keep using SGD to optimize the hyper-parameter C on the upper level.

The bilevel hyper-parameter optimization problem we are solving becomes:

$$\begin{aligned} \underset{C,\bar{\boldsymbol{\alpha}}}{Min}\ F(C,\bar{\boldsymbol{\alpha}}) \ &= \sum_{j=1}^{N} \max\{0, 1 - y_j^v\left(\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{x}_j^v\right)\} \\ &= \sum_{j=1}^{N} \max\{0, 1 - y_j^v\left(\left(\sum_{i=1}^{L} \bar{\alpha}_i y_i x_i\right)^{\boldsymbol{T}}\boldsymbol{x}_j^v\right)\} \\ &= \sum_{j=1}^{N} \max\{0, 1 - y_j^v\left((\boldsymbol{M}\boldsymbol{\alpha})^{\boldsymbol{T}}\boldsymbol{x}_j^v\right)\} \end{aligned} \tag{5.22}$$

$$\begin{aligned} subject\ to \quad & C_{\max} \geq C \geq 0 \\ & \bar{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha}}{argmin}\left\{f(\boldsymbol{\alpha}) = \boldsymbol{e}^T\boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}^T Q\boldsymbol{\alpha} : 0 \leq \alpha_i \leq C\right\}, \forall\, i \in \{1, \ldots, L\} \end{aligned}$$

where $Q_{ij} = y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j$ for the training set and $\boldsymbol{e}$ is a unit vector, $M_{*,j} = y_j \boldsymbol{x_j}$ (in training set).

Using chain rule, we have:

$$\nabla_C F\left(C, \boldsymbol{\alpha}(C)\right) = \nabla_C F(C, \boldsymbol{\alpha})^T + \nabla_{\boldsymbol{\alpha}} F\left(C, \boldsymbol{\alpha}(C)\right)^T J_{\boldsymbol{\alpha}}(C) \tag{5.23}$$

Following the framework by Couellan and Wang (2015), stochastic gradient descent can be used to obtain a noisy unbiased estimate of the gradient by randomly choosing one data sample q in $F(C, \boldsymbol{\alpha})$.

$$\nabla_C F_q\big(C, \boldsymbol{\alpha}(C)\big) = \nabla_C F_q(C, \boldsymbol{\alpha})^T + \nabla_{\boldsymbol{\alpha}} F_q\left(C, \boldsymbol{\alpha}(C)\right)^T J_{\boldsymbol{\alpha}}(C) \tag{5.24}$$

From Eq. (5.22), we have $\nabla_C F_q(C, \boldsymbol{\alpha})^T = 0$ and $\nabla_{\boldsymbol{\alpha}} F_q\left(C, \boldsymbol{\alpha}(C)\right)^T = -y_q^v\big(\boldsymbol{x}_q^v\big)^T M$.

Therefore, Eq. (5.24) becomes:

$$\nabla_C F_q\big(C, \boldsymbol{\alpha}(C)\big) = -y_q^v\big(\boldsymbol{x}_q^v\big)^T M J_{\boldsymbol{\alpha}}(C) \tag{5.25}$$

The key is how to obtain $J_{\boldsymbol{\alpha}}(C)$. As we use the dual coordinate descent method by Hsieh et al. (2008) to solve the lower level problem [12], the optimality condition for the lower problem is that the projected gradient is zero, i.e., $\nabla^P f(\bar{\boldsymbol{\alpha}}) = 0$, where $f(\boldsymbol{\alpha}) = \boldsymbol{e}^T\boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}^T Q\boldsymbol{\alpha}$. According to Hsieh et al. (2008) [12], we have:

$$\nabla_i^P f(\boldsymbol{\alpha}) = \begin{cases} \nabla_i f(\boldsymbol{\alpha}) & if\ 0 < \alpha_i < C \\ \min\big(0, \nabla_i f(\boldsymbol{\alpha})\big) & if\ \alpha_i = 0 \\ \min\big(0, \nabla_i f(\boldsymbol{\alpha})\big) & if\ \alpha_i = C \end{cases} \tag{5.26}$$

The only α_i that are a function of C are the ones that are equal to C, i.e., the constraint $\alpha_i \leq C$ is active when optimality condition is satisfied. Therefore,

$$J_{\boldsymbol{\alpha}}(C)_i = \begin{cases} 1, & if\ \alpha_i = C \\ 0, & if\ 0 < \alpha_i < C \end{cases} \tag{5.27}$$

The advantage of our approach is that we didn't introduce any additional hyper-parameters to the lower level optimization problem by using the DCD method [12]. As the DCD method is very efficient for large-scale SVM problems, we believe the new approach to be computational fast, while achieving the optimal solution for hyper-parameter C.

The specific algorithm for our approach is shown in Algorithm 1 below. As we combined SGD with DCD in Algorithm 1, we also call Algorithm 1 the SGD+DCD method in the following section.

Algorithm 1: Stochastic gradient descent combined with dual coordinate descent method for Linear SVM, SGD+DCD

Set C_{min}, C_{max}, initialize C, $\boldsymbol{\alpha}$ and the corresponding $\boldsymbol{\beta} = \sum_1^L \alpha_i y_i \boldsymbol{x_i}$, $M_{*,j} = y_j \boldsymbol{x}_j$;
while *Stopping criteria not satisfied* **do**
  Pick l randomly from training set $T_e = \{1, ..., L\}$;
  if $\boldsymbol{\alpha}$ *is not optimal* **then**
    $\bar{\alpha}_l \leftarrow \alpha_l$;
    $\nabla_l f(\boldsymbol{\alpha}) = y_l \boldsymbol{\beta}^T \boldsymbol{x_l} - 1$;
    $$\nabla_l^P f(\boldsymbol{\alpha}) = \begin{cases} min(\nabla_l f(\boldsymbol{\alpha}), 0) & \text{if } \alpha_l = 0 \\ max(\nabla_l f(\boldsymbol{\alpha}), 0) & \text{if } \alpha_l = C \\ \nabla_l f(\boldsymbol{\alpha}) & \text{if } 0 < \alpha_l < C \end{cases} ;$$
    if $|\nabla_l^P f(\boldsymbol{\alpha})| \neq 0$ **then**
      $\alpha_l \leftarrow min(max(\alpha_l - \nabla_l f(\boldsymbol{\alpha})/Q_{ll}, 0), C)$;
      $\boldsymbol{\beta} \leftarrow \boldsymbol{\beta} + (\alpha_l - \bar{\alpha}_l) y_l \boldsymbol{x_l}$;
  $$J_{\boldsymbol{\alpha}}(C)_i = \begin{cases} 1, & \text{if } \alpha_i = C \\ 0, & \text{if } 0 \leq \alpha_i < C \end{cases}$$
  if $J_{\boldsymbol{\alpha}}(C)$ *are not all zeros* **then**
    Pick q randomly from validation set $V_e = \{j = 1, ..., N | y_j^v \boldsymbol{\beta}^T \boldsymbol{x_j}^v < 1\}$;
    Compute $\nabla_{\boldsymbol{\alpha}} F_q(C, \boldsymbol{\alpha}(C)) = -y_q^v (\boldsymbol{x}_q^v)^T M J_{\boldsymbol{\alpha}}(C)$;
    $C \leftarrow C - \eta_C \nabla_{\boldsymbol{\alpha}} F_q(C, \boldsymbol{\alpha}(C))$;
    if $C < C_{min}$ **then**
      $C = C_{min}$;
    if $C > C_{max}$ **then**
      $C = C_{max}$;

5.3.4 *K*-fold Cross-validation

In practice, k-fold cross-validation is the standard approach to tune hyper-parameters [119]. We can easily extend the one-fold validation problem formulation in the previous section to the k-fold cross-validation case.

Let's assume we randomly split the dataset into k folds. Denote the data in the k^{th} validation fold as $\left(\boldsymbol{x}_i^{v_k}, y_i^{v_k}\right), i \in \{1,2,\dots N\}$, where N is the number of samples in the validation fold. The rest of the dataset, i.e., $k-1$ folds are the training data. Let's denote the training data as $\left(\boldsymbol{x}_i^{t_k}, y_i^{t_k}\right), i \in \{1,2,\dots L\}$, where L is the number of samples in the training data. Figure 5.2 illustrates the data partition for 3-fold cross-validation.

The objective function on the upper level becomes minimizing the average loss over the k validation folds. On the lower level, we are minimizing the loss for each of the k corresponding training data. The bilevel hyper-parameter optimization problem becomes:

$$\begin{aligned}
&\underset{C,\bar{\boldsymbol{\alpha}}_k}{Min}\, F(C,\bar{\boldsymbol{\alpha}}_k) = \frac{1}{K}\sum_{k=1}^{K}\sum_{j=1}^{N}\max\{0, 1 - y_j^{v_k}\left((\boldsymbol{M}_k\bar{\boldsymbol{\alpha}}_k)^T\boldsymbol{x}_j^{v_k}\right)\} \\
&= \frac{1}{K}\sum_{k=1}^{K}\sum_{j=1}^{N_{ek}}(1 - y_j^{v_k}\left((\boldsymbol{M}_k\bar{\boldsymbol{\alpha}}_k)^T\boldsymbol{x}_j^{v_k}\right)) \qquad (5.28) \\
subject\ to\quad & C_{\max} \geq C \geq 0 \\
& \bar{\boldsymbol{\alpha}}_k = \underset{\boldsymbol{\alpha}_k}{argmin}\left\{f(\boldsymbol{\alpha}_k) = \boldsymbol{e}^T\boldsymbol{\alpha}_k - \frac{1}{2}\boldsymbol{\alpha}_k^T Q_k \boldsymbol{\alpha}_k : 0 \leq \alpha_{k_i} \leq C\right\} \\
& \forall\, i \in \{1,\dots,L\}, k \in \{1,\dots,K\}
\end{aligned}$$

where $Q_{k_{ij}} = y_i{}^{t_k} y_j{}^{t_k} \boldsymbol{x}_i^{t_k T}\boldsymbol{x}_j^{t_k}$, $M_{k_{*,j}} = y_j{}^{t_k}\boldsymbol{x}_j^{t_k}$ and $\boldsymbol{e}$ is a unit vector.

Given the value of C, all the lower-level optimization problems are independent. Therefore, we can update the variable $\boldsymbol{\alpha}_k$ for each lower variable independently. The only modification of the

SGD+DCD method needed to perform K-sfold cross-validation is at step 15 in Algorithm 1. Here we update C using the averaged gradient from all the training folds instead, i.e., using $\frac{1}{K}\sum_{k=1}^{K}\nabla_{\boldsymbol{\alpha}_k}F_p\left(C,\boldsymbol{\alpha}_k(C)\right)$.

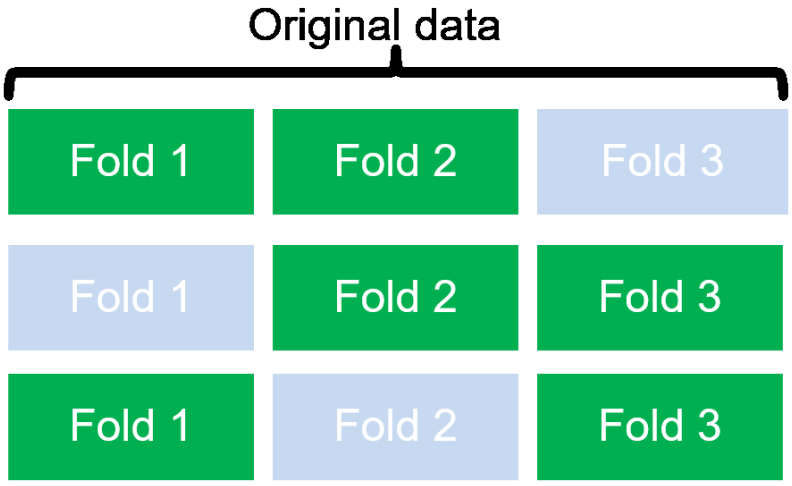


Figure 5.2 Data partition for 3-fold cross-validation. Each row represents a split of the original data into training and validation fold. Training fold is in green. The corresponding validation fold is in blue.

Next, we are going to illustrate the effectiveness of our approach with numerical results on multiple benchmark datasets. All the experiments were conducted using 5-fold cross-validation.

5.4 Numerical Experiment

There are two main implementation issues that need to be clarified. First, a stopping criterion needs to be chosen. In practice, several stopping criteria are used. For instance, stop the algorithm when the accuracy exceeds a predefined threshold. Or simply stop the algorithm when the maximal number of iterations is reached. To compare the two methods, we used the second stopping criterion.

Another implementation issue is how to initialize all the variables. For our SGD+DCD method, we initialized $\boldsymbol{\alpha}$ as zero vector as suggested by [12] because the solution of $\boldsymbol{\alpha}$ typically has only a few nonzero elements, which are the support vectors. For the bilevel-SGD method, we initialized all the variables following Couellan and Wang's (2015) study [131]. Specifically, we randomly initialized the $\boldsymbol{\beta}$ from a $[0,1]$ uniform distribution. The learning rate for the lower level SGD was chosen to be $\frac{1}{t}$, and the learning rate for the upper level SGD was set as $\frac{1}{t\sqrt{p}|\nabla_C F_q(C,\boldsymbol{\beta}(C))|}$ where t is the t^{th} iteration and p is the number of features. Our SGD+DCD method used $\frac{1}{t\sqrt{p}|\nabla_C F_q(C,\boldsymbol{\alpha}(C))|}$ as the learning rate η_C for the upper level SGD. For both methods, the value for hyper-parameter C was initialized to be 10^{-4}. In addition, we set $C_{\min} = 10^{-4}$, and $C_{max} = 10^{6}$ for both methods.

The datasets we used are the following:

1. Cancer: Wisconsin diagnostic breast cancer dataset. The two classes are the benign or malignant diagnosis. Number of samples: 699, number of features 9. Source: UCI Machine Learning repository [136].

2. Pima: Diabetes data. The two classes are positive or negative diabetes diagnosis. Number of samples: 768, number of features: 8. Source: UCI Machine Learning repository [137].

3. SVMguide1: astroparticle dataset. Number of samples: 3089, number of features: 4. Source: LIBSVM – a Library for Support Vector Machines [138].

4. Connect: connect-4 game dataset. The prediction target is win or loss. Number of samples: 67557, number of features: 42. Source: UCI Machine Learning repository [139].

5. Magic04: gamma telescope dataset. The prediction target is whether a signal is high energy gamma signal or background signal. Number of samples: 19020, number of features: 10. Source: UCI Machine Learning Repository [140].

6. Xerostomia: head and neck cancer radiotherapy side effect dataset. The prediction target is whether a patient develops acute xerostomia after radiotherapy. Number of samples: 551, number of features: 943. A detailed description of this dataset is in Chapter 3 of this thesis. Source: Oncospace JHU [141].

7. Xerostomia recovery: head and neck cancer radiotherapy side effect dataset. The prediction target is whether a patient developed xerostomia recovered at 18 months post radiotherapy. Number of samples: 146, number of features: 943. A detailed description of this dataset is in Chapter 3 of this thesis. Source: Oncospace JHU [141].

All dataset features were preprocessed and scaled to be in the range of [-1,1]. The class labels were coded as $y \in \{-1,1\}$. These three datasets were used by Couellan and Wang's (2015) study. We used them here to compare the two methods. For 5-fold cross-validation, we created the five folds using stratified sampling to ensure each fold has the same distribution of the prediction labels.

Table 5.1 reports the numerical results of the two methods. Accuracy is the prediction accuracy on the validation folds (hold-out data) by doing 5-fold cross-validation. C is the final value obtained for the hyper-parameter. Running time is the total CPU (Intel Core i7-7700HQ) time spent on training the SVM model. We obtained the results presented in Table 5.1 by running each of the methods for 150 iterations.

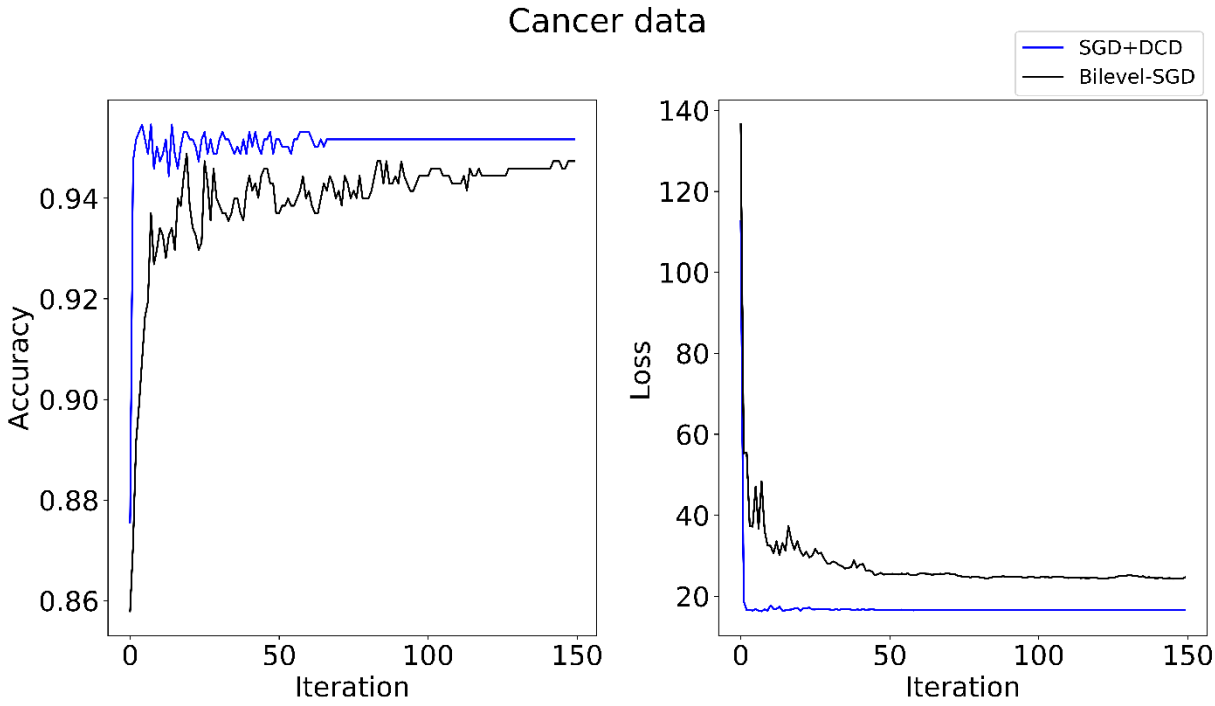
Table 5.1 Numerical results of the bilevel-SGD method and the SGD+DCD method.

Dataset	**Bilevel-SGD**			**SGD+DCD (our approach)**		
	Accuracy (%)	C	**Running time** (s)	**Accuracy** (%)	C	**Running time** (s)
Cancer	94.29	1.24	0.30	95.02	0.16	0.49

Pima	65.58	0.38	0.33	76.53	0.21	0.91
SVMguide1	72.81	1.06	1.21	84.88	0.29	2.43
Connect	72.77	0.0001	89.36	83.68	0.06	148.39
Magic04	66.97	0.26	22.80	78.55	0.54	47.39
Xerostomia	68.97	0.10	2.56	69.51	0.01	2.57
Xerostomia recovery	81.58	0.12	0.39	78.75	0.01	0.45

From Table 5.1, it's easy to see with a slight increase of running time, the SGD+DCD method achieved higher accuracy on the validation folds for all datasets except the xerostomia recovery dataset. For the Cancer dataset, xerostomia and xerostomia recovery dataset, the accuracy result is about the same. For all the other datasets, the SGD+DCD methods accuracies are more than 10% higher than the standard bilevel-SGD method. For high dimensional datasets, xerostomia and xerostomia recovery, the SGD+DCD has about the same running time with the bilevel-SGD method. For the other non-high dimensional datasets, the bilevel-SGD is relatively faster.

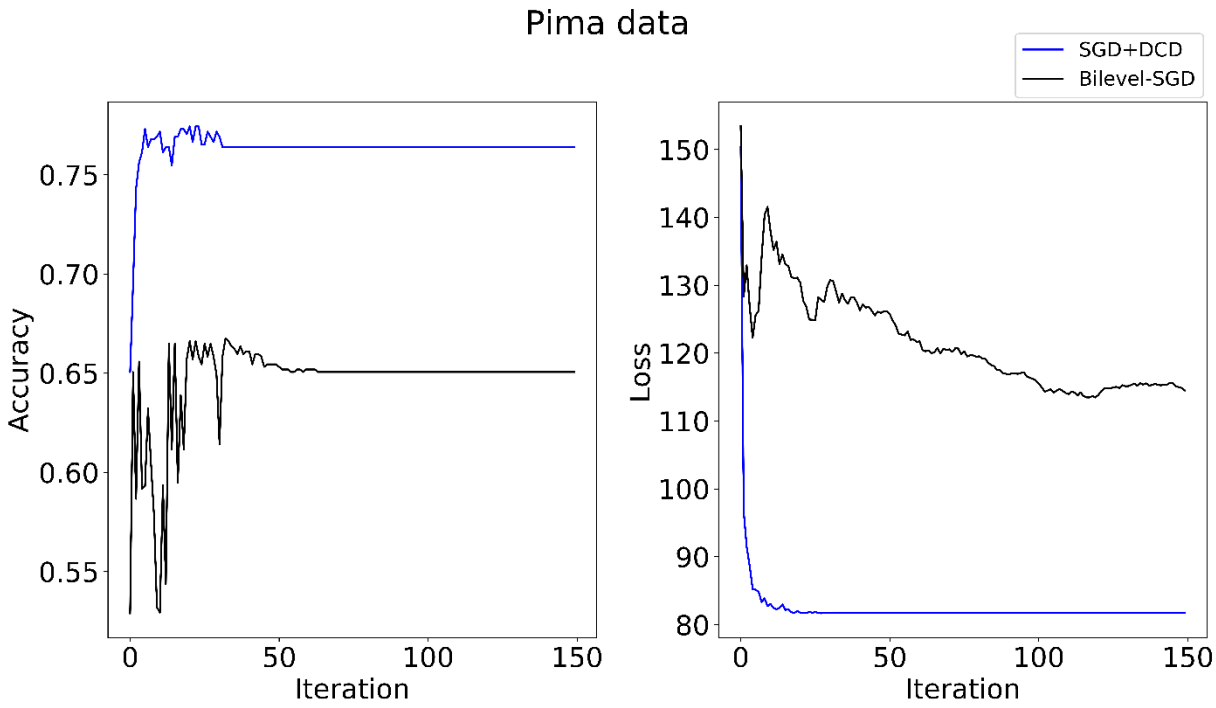
To further compare the convergence results, we plotted the accuracy and loss profile on seven datasets for both methods in Figure 5.3 to Figure 5.9. We can see that the SGD+DCD method converges to the optimal solution much faster than the bilevel-SGD method. Also, the SGD+DCD has higher accuracy and a lower loss on the validation fold than the bilevel-SGD method for all three datasets.



(a) Accuracy

(b) Loss on the validation set

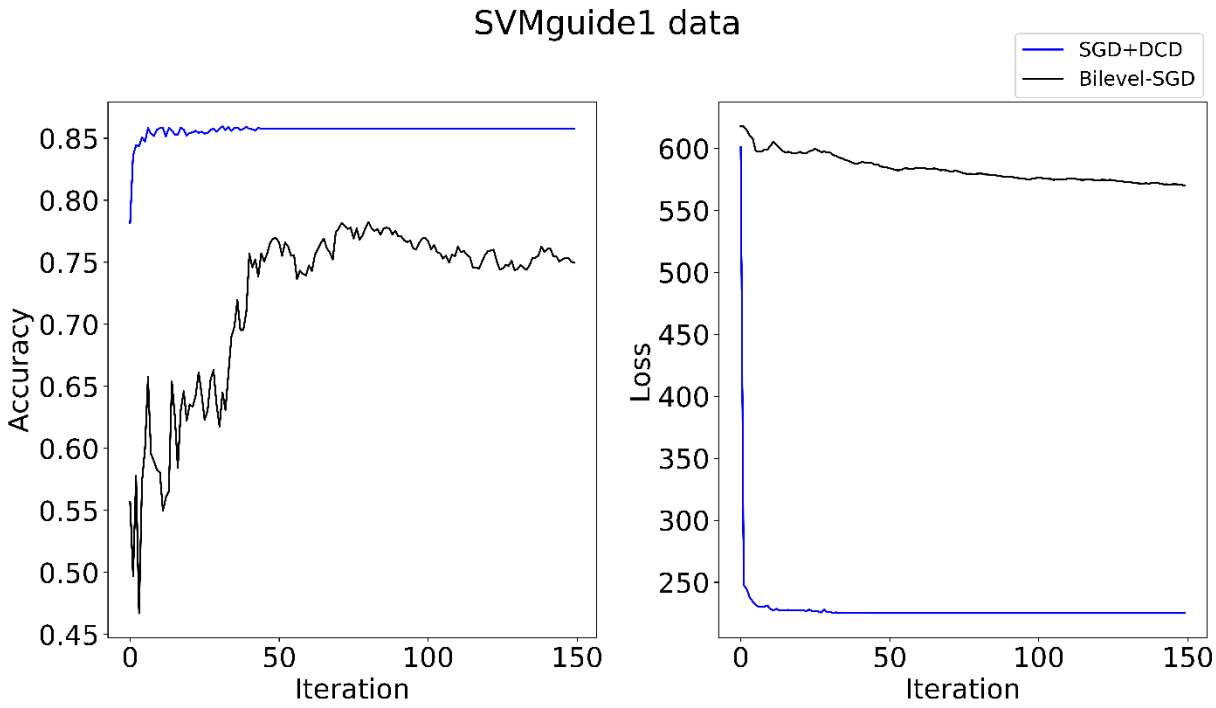
Figure 5.3 The numerical results of running Algorithm 1 (our SGD+DCD) method and the bilevel-SGD method on the Cancer dataset. (a) shows the prediction accuracy on the validation folds as the algorithm proceeds; (b) shows the loss on the validation folds as the algorithm proceeds.



(a) Accuracy

(b) Loss on the validation set

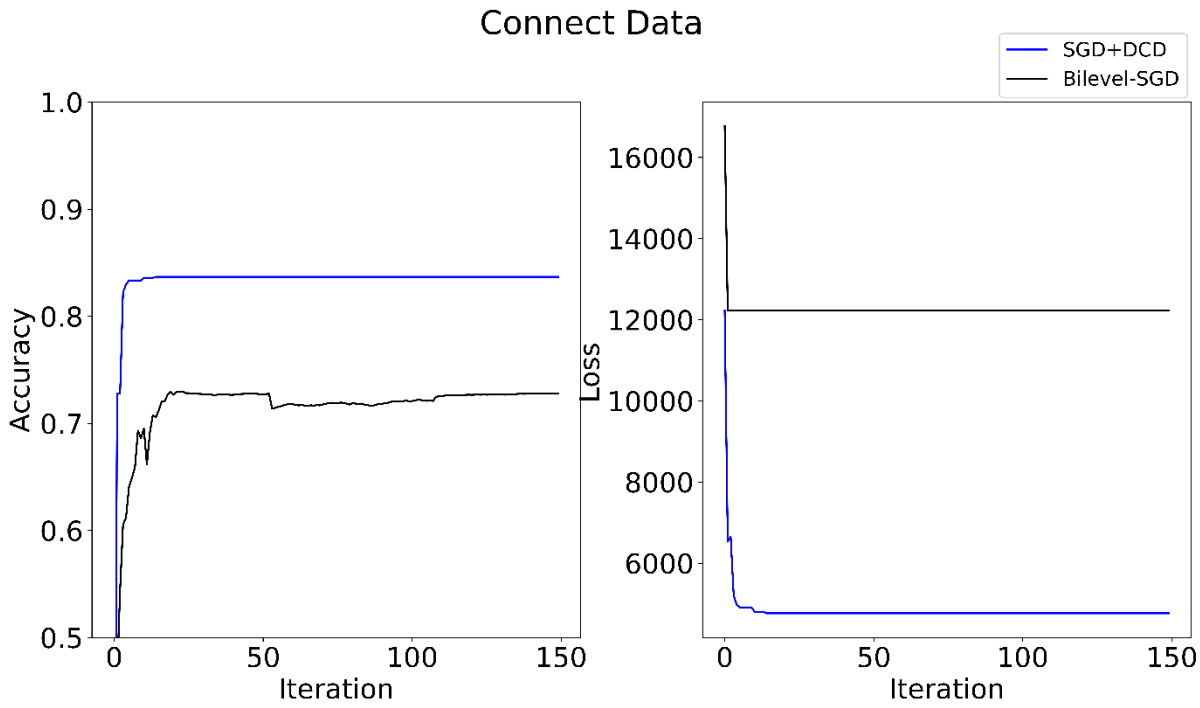
Figure 5.4 The numerical results of running Algorithm 1 (our SGD+DCD) method and the bilevel-SGD method on the Pima dataset. (a) shows the prediction accuracy on the validation folds as the algorithm proceeds; (b) shows the loss on the validation folds as the algorithm proceeds.



(a) Accuracy

(b) Loss on the validation set

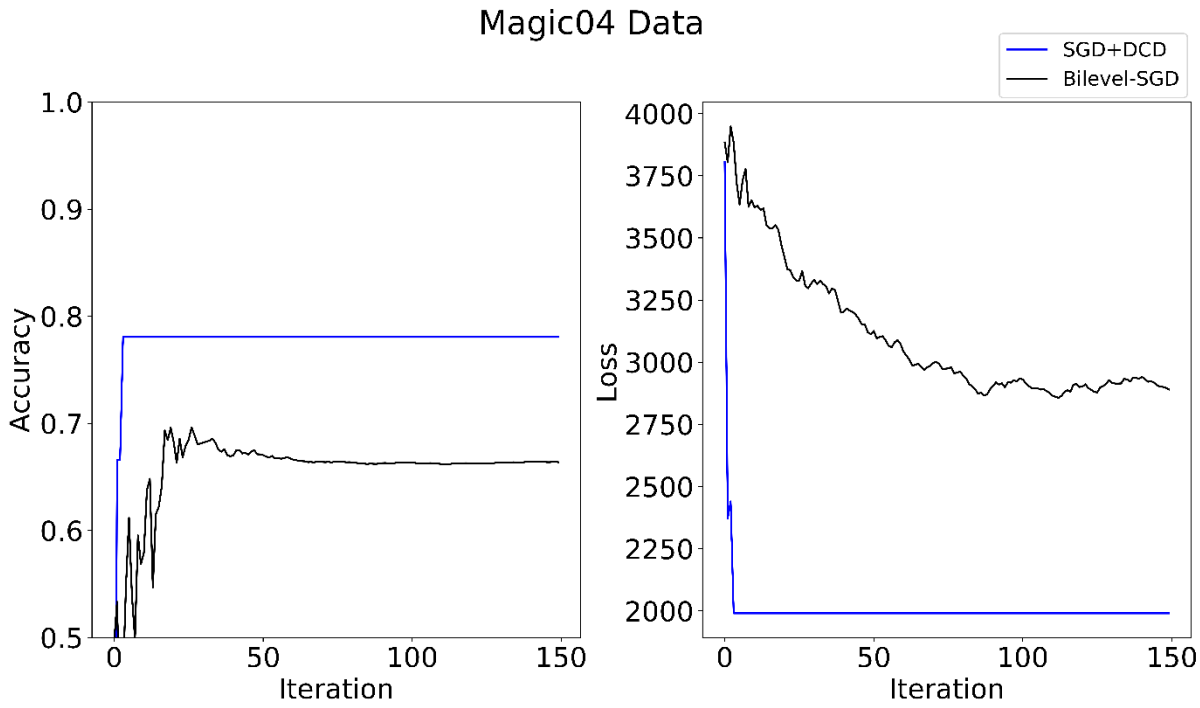
Figure 5.5 The numerical results of running Algorithm 1 (our SGD+DCD) method and the bilevel-SGD method on the SVMguide1 dataset. (a) shows the prediction accuracy on the validation folds as the algorithm proceeds; (b) shows the loss on the validation folds as the algorithm proceeds.



(a) Accuracy

(b) Loss on the validation set

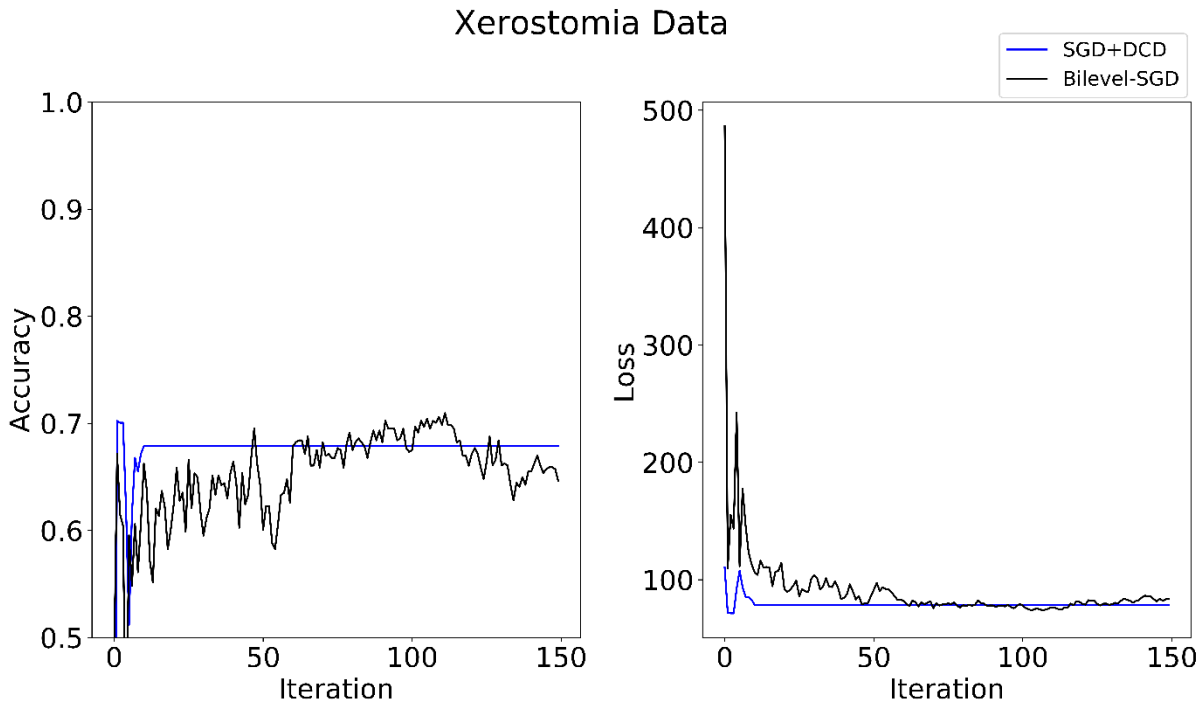
Figure 5.6 The numerical results of running Algorithm 1 (our SGD+DCD) method and the bilevel-SGD method on the Connect dataset. (a) shows the prediction accuracy on the validation folds as the algorithm proceeds; (b) shows the loss on the validation folds as the algorithm proceeds.



(a) Accuracy

(b) Loss on the validation set

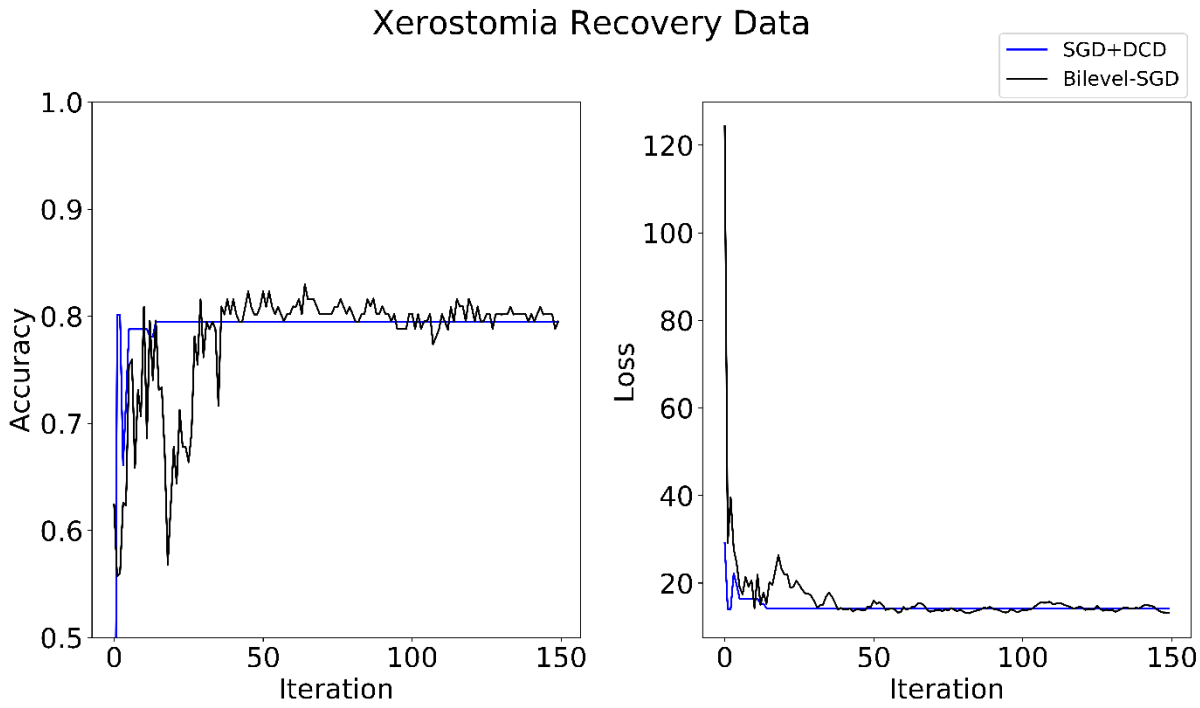
Figure 5.7 The numerical results of running Algorithm 1 (our SGD+DCD) method and the bilevel-SGD method on the Magic04 dataset. (a) shows the prediction accuracy on the validation folds as the algorithm proceeds; (b) shows the loss on the validation folds as the algorithm proceeds.



(a) Accuracy

(b) Loss on the validation set

Figure 5.8 The numerical results of running Algorithm 1 (our SGD+DCD) method and the bilevel-SGD method on the Xerostomia dataset. (a) shows the prediction accuracy on the validation folds as the algorithm proceeds; (b) shows the loss on the validation folds as the algorithm proceeds.



(a) Accuracy

(b) Loss on the validation set

Figure 5.9 The numerical results of running Algorithm 1 (our SGD+DCD) method and the bilevel-SGD method on the Xerostomia Recovery dataset. (a) shows the prediction accuracy on the validation folds as the algorithm proceeds; (b) shows the loss on the validation folds as the algorithm proceeds.

5.5 Discussion and Future Work

Even though the SGD+DCD converges to an optimal solution very fast, it's computationally more expensive than the bilevel-SGD method. Bilevel-SGD is faster because it uses SGD to optimize

both the lower and upper-level problems. As we know, SGD is fast due to cheap gradient updates. For our SGD+DCD method, we optimized the lower-level problem using DCD. DCD method iterates through all the variables on the lower level for each outer iteration. For DCD, the number of variables on the lower-level equals the number of training examples. Therefore, when the number of training examples is much larger than the number of features, DCD requires longer running time [12]. All the datasets we used (except the Xerostomia and Xerostomia Recovery data) are all low dimensional datasets. The largest number of features among the seven datasets except the xerostomia data is 42 (the Connect data), which is much smaller than its number of samples (699). DCD is well suited for high dimensional dataset where the number of features is larger than the number of samples [12]. As a result, we believe the SGD+DCD method will be computationally more efficient to tune the hyper-parameter for high dimensional datasets, as shown by the results on the high-dimensional Xerostomia and Xerostomia Recovery datasets.

There are two main reasons that our SGD+DCD method achieves better accuracy performance than the bilevel-SGD method. Using SGD for optimizing both the hyper-parameter C on the upper level and SVM parameters on the lower level is fast due to cheap gradient estimates. However, there are assumptions and approximations behind the bilevel-SGD approach, which enables our SGD+DCD method to achieve better accuracy performance. The first reason is that, for each iteration of updating hyper-parameter C, we are assuming the lower-level SVM problem has already reached its optimal solution. This is not always true for the bilevel-SGD method,

especially in the early iterations when the lower-level SVM problem is far from converging to the optimal solution. Regarding this assumption, our SGD+DCD method is more appropriate than Couellan and Wang (2015)'s bilevel-SGD method as the DCD method converges much faster than the SGD method for solving the lower-level SVM problem.

The second reason is that Couellan and Wang (2015)'s method approximated the optimality condition of the lower-level problem using only one randomly sampled data instance when using SGD to optimize *C*. Specifically, they estimated the stochastic gradient using Eq. (5.17), which is an approximation of the true optimality condition (Eq. (5.16)). They computed the stochastic gradient using the implicit differentiation trick on the optimality condition. This is different from a regular SGD problem, where the stochastic gradient is usually computed from a loss function. We utilized the implicit differentiation trick as well to obtain the gradient of lower-level SVM parameter $\boldsymbol{\alpha}$ with respect to C, i.e., $J_{\boldsymbol{\alpha}}(C)$. However, to obtain $J_{\boldsymbol{\alpha}}(C)$, we didn't approximate the optimality condition of the lower-level SVM problem. Instead, we computed the exact gradient of $J_{\boldsymbol{\alpha}}(C)$ using Eq. (5.26) and Eq. (5.27). This is both an exact and computationally cheap gradient calculation as most of the elements in $J_{\boldsymbol{\alpha}}(C)_i$ are zero when the lower-level SVM problem reaches its optimal solution.

In summary, we developed an efficient hyper-parameter optimization method to optimize the regularization parameter C for supoort vector machine by combining stochastic gradient descent and dual coordinate descent method. Several benchmark datasets show that our method

yields consistently higher out-of-sample accuracy performance than an existing bilevel-SGD method with slightly increased running time.

Chapter 6

Conclusion and Future Work

Through the last four chapters, we illustrated how we solved complex healthcare and energy systems problems utilizing machine learning and optimization techniques. As information technology advances, accumulating healthcare and energy data has revealed many critical problems we are facing in these systems. Together with the recent rapid development in machine learning and optimization modeling methods, previously intractable problems now become solvable. For instance, hospital readmission data reveals that heart failure patients have one of the highest 30-day readmission rates. By better managing hospital readmissions, we can not only reduce our ever-increasing healthcare costs but also improve patients' treatment outcome and quality of life. The daily risk prediction model can help the physicians target patients with high readmission risk and provide targeted interventions to improve the patients' condition. In the

radiation oncology department, patients' quality of life is often compromised due to radiation-induced side effects. Oncospace is an analytic relational database that centers on an informatics infrastructure established at the Johns Hopkins hospital in 2008. This database has been systematically capturing patients' data at all phases of their care. Using prospectively collected patients' assessments and treatment data, we were able to discover new knowledge about radiation dose effect on xerostomia. This new knowledge can help physicians and researchers better understand the sensitivity of salivary function given radiation dose.

For energy systems, we were able to perform scenario analysis on potential US and EU's renewable energy policies about wood chip using equilibrium modeling. Effective and sustainable renewable energy policies are crucial to reducing carbon emissions in order to mitigate climate change. Renewable energy policies are often created within a local region, such as within a country or a state. However, the scale of its impact on environmental sustainability can be global due to international trade of bioenergy products. We identified potential detrimental global sustainability issues with the policy in which the US is considering to include wood chip as a renewable energy source. Precisely, according to our model and historical wood chip trade data, the policy is likely to cause deforestation in countries in Southeast Asia, Latin America, and the Former Soviet Union.

Finally, through Chapter 5, we have seen how optimization techniques can be used to improve the machine learning algorithms regarding hyper-parameter tuning. A more systematic and efficient approach for building machine learning models and tuning hyper-parameters is

needed to help us move away from ad hoc empirical approaches. Tuning hyper-parameters for deep learning models has been notably called art, not science. Gradient-based optimization techniques for hyper-parameter optimization have been emerging recently with the advances of deep learning. We demonstrated that applying the same framework but adapted to exploit the special structure of a classical machine learning algorithm, SVM, enables us to tune the hyper-parameter for SVM efficiently.

Despite the contributions and advances we have made to solve those problems in healthcare and energy systems, certain limitations exist. To overcome these limitations and ultimately solve those problems, I will discuss future directions of research for the projects described in Chapter 2 to Chapter 5 next.

6.1 Holistic Approach to Reducing Hospital Readmission

Hospital readmission is a complex issue. Many factors can lead to hospital readmissions, not mention the uncertainties inherited in each of those factors. For instance, 30-day readmission for heart failure patients may be caused by being discharged with an unstable condition. Or patients tend to revisit hospitals more frequently because they live close to the hospitals, while patients who live further away from the medical facility avoid readmissions due to the geographical distances. Or patients who live with their family members could receive better care and tend to revisit hospital more often than patients who live alone. All those clinical, geographical, and

societal factors could affect the readmission rates. A holistic approach may be necessary to reduce hospital readmission rates. A holistic approach involves improving patients' treatment during hospital stays and also patients-care post hospital discharges. It also requires a more complete data capturing process. Currently, the data available for studying hospital readmissions are mostly patients' data collected during patients' hospital stays. The data about patients' care and health status post-discharge is often unknown. However, we believe that patients' health status and care information post discharge may be more important for reducing readmissions. To collect patients' data post-discharge is a challenging issue. First, we need the informatics infrastructures built out of the hospital. Second, closely monitoring patients' health status post-discharge may encounter legal privacy issues.

Finally, patients' electronic health records can be shared across all the hospitals the patients visited. One challenge in our heart failure patient readmission prediction work is that patients sometimes were transferred to other hospitals. As a result, patients effectively dropped out from the study as our study hospital only has access to patients' data within that hospital. Therefore, an informatics infrastructure that stores all the historical healthcare data for a patient across different hospitals will be extremely valuable. First, a complete historical healthcare dataset for each patient enables more accurate research results and findings, which leads to better understanding and treatment of diseases. Second, a complete healthcare record provides hospital more information about the patient, which enables a more precise personal medical treatment.

Regarding modeling patients' readmission risk, a dynamic and multi-module prediction model that combines different types of patients' information may be developed. A multi-module prediction model utilizes different modules of data and treats them separately. For instance, one module can be patients' longitudinal laboratory test data. Another module is patients' geographical, societal information. Patients' treatment information can also be an important module. That information has different characteristics and can give different indications on patients' health condition and readmission risk. Separate models can be built for predicting readmission risk using each module's data. Finally, these models can be integrated into a multi-module prediction model.

In summary, we believe this holistic approach can help us successfully manage hospital readmissions, improve patients' quality of life, and save healthcare cost.

6.2 Optimal Treatment Planning Considering Toxicity in Radiation Oncology

For this radiation oncology work, there are different goals. The ultimate objective is to optimize the radiation treatment plans for a head and neck cancer patients. A radiation treatment plan is considered optimal if it minimizes patients' treatment side effects as well as maximizes its treatment effect of killing cancer cells. As we can see, this is a multi-objective optimization problem, and a single optimal solution may not exist. Likely, there is a tradeoff between two

objectives. Higher radiation dose kills more cancer cells but leads to severe treatment side effects. Lower radiation dose can reduce treatment side effects, but also compromises the effect of killing cancer cells. We believe, a particular approach is to, first, predefine a treatment effect requirement on killing cancer cells. Given this requirement, we can optimize the radiation treatment plan to minimize the side effect for a patient. Essentially, we can reduce the multi-objective optimization problem into a single objective optimization problem with the requirement of killing cancer cells as constraints.

To optimize treatment plan, another objective of this work is to understand how radiation therapy affects treatment side effects. Given the complex organ structures within head and neck, understanding this relationship is a difficult task, and it's a popular ongoing research area. Understanding how radiation dose in different subvolumes affects side effects will enable us to optimize the treatment plans to minimize those side effects. For instance, if we know a region within the parotid gland is very sensitive to radiation but very important in preserving salivary function, we can try to avoid radiating this particular region during treatment planning.

Currently, we used a regularized logistic regression model to study the spatial relationship between dose and xerostomia. The challenge is that our findings are limited by the dose variation within the patient cohort. Our current approach is not able to identify a causal relationship between spatial dose and xerostomia. Ideally, a clinical trial can be conducted to study the causal relationship and the biological mechanism behind it.

Another limitation of the current approach is that we treated each voxel dose features as an independent feature, without modeling the spatial dependencies between these individual dose features. We believe a model that captures the spatial dependencies between voxel dose features will improve our understanding of the spatial dose effect on side effects. It could also improve the prediction performance for predicting side effects. A convolutional neural network (CNN) is a supervised machine learning model that has been successfully used for image classification and medical image segmentation tasks [142]. By convoluting over the spatial input data using multiple sliding windows (filters), CNN can automatically learn spatial features in the input data. For instance, CNN can learn edges and various objects automatically from image data. Our voxel dose data is very similar to image dose. The only difference is that the voxel dose data is 3-dimensional, but the same technique of CNN applies to the 3-dimensional dataset. We can use a 3-dimensional sliding window to learn the spatial features. We have built a CNN model to use our 3-dimensional voxel dose data to predict xerostomia and currently optimize the model structure to improve its prediction performance. Once we have a CNN model with good predictive performance, we can use it for automatic feature discovery as opposed to performing feature engineering manually.

6.2.1 Incorporating Toxicity Outcomes in Treatment Planning Optimization

The goal for radiation treatment planning optimization is maximizing the dose delivered to tumor cells and minimizing the dose delivered to the surrounding normal organs and tissues. The original optimization problem can be formulated as a multi-objective optimization problem [143]:

$$\underset{\boldsymbol{x}}{Min}\, F(\boldsymbol{x}) = \alpha_1 f_1(\boldsymbol{x}) + \cdots + \alpha_{\boldsymbol{n}} f_n(\boldsymbol{x}) \tag{6.1}$$

$$subject\ to \quad \boldsymbol{x} \in \mathsf{C}$$

where $\boldsymbol{x}$ is the planned radiation dose, and α are the weights assigned to different objectives.

To explicitly incorporating toxicity outcome into the optimization problem, we can add the predicted risk of developing certain toxicity, for instance, xerostomia, to the objective function of the original optimization problem as another penalty term. The predicted risk is a function of the radiation dose $\boldsymbol{x}$ estimated by our trained risk prediction model.

In our case, we will use the trained ridge logistic regression model, instead of trained lasso logistic regression to predict the risk of toxicity. The reason is that trained lasso logistic regression model yields a sparse solution where only a small subset of the voxel dose features can be selected as the final features. If we use this sparse subset of voxel dose features as the decision variables in the optimization model, we will be only optimizing the radiation dose in that subset of voxels and wouldn't be able to optimize the radiation dose in all the other regions, which is not applicable for our problem. On the other hand, ridge logistic regression doesn't yield a sparse solution and assigns a non-zero weight for most of the voxel-dose features. Therefore, using the trained ridge logistic regression model, we would be able to optimize the radiation dose in all regions while minimizing its effect on predicted toxicity outcomes.

Precisely, the new optimization problem explicitly incorporating predicted risk of toxicity using ridge logistic regression model is as follows:

$$\underset{x}{Min}\ G(x) = \gamma_1 F(\boldsymbol{x}) + \gamma_2 \frac{1}{1 + e^{-\boldsymbol{\beta}^T \boldsymbol{x}}} \qquad (6.2)$$

$$subject\ to \quad \boldsymbol{x} \in \mathbb{C}$$

where $\frac{1}{1+e^{-\boldsymbol{\beta}^T \boldsymbol{x}}}$ is the predicted risk of toxicity using the trained ridge logistic regression model, γ_1 and γ_2 are the weights assigned to the original objective and the predicted risk of toxicity, and $\boldsymbol{x}$ is the voxel-based radiation dose.

6.3 Modeling Wood Chip Trade as a Biofuel

The main challenge for this work is that there is not sufficient data available to distinguish wood chip traded for paper production versus for biofuel production. Most of the wood chip has been traded for producing papers historically. Renewable energy policies have increased the demand for wood chip as biofuel. To more realistically model the effect of renewable energy policies on wood chip trade, we should start recording trade data separately for wood chip traded for paper production and biofuel production. In other words, we should refine our trade data recording based on the final usage of the commodity. Given wood chip trade data for different usages, we can build a trade model which models wood chip for paper and biofuel production separately. Also, we can model the interaction or substitution effect between the two types of wood chip. We believe this approach will lead to a more realistic policy scenario analysis on wood chip.

6.4 Automatic Hyper-parameter Tuning for Machine Learning Models

Emerging research on creating gradient-based methods for hyper-parameter tuning for machine learning models is expected to create machine learning algorithms that can tune hyper-parameters automatically. All of those methods rely on the implicit differentiation trick. The implicit differentiation trick enables us to obtain the gradient estimate for the hyper-parameters, which is the key for these methods. However, this approach has a large memory requirement, especially for deep learning models. We believe that, for future work, we can develop new ways to estimate the gradient for hyper-parameters that is memory efficient. Another option would be to keep using the implicit differentiation trick but exploring new techniques to overcome the high computational burden [123].

As for hyper-parameter optimization for SVM, all current methods almost exclusively only optimized the regularization hyper-parameter C. No efficient methods exist for optimizing the nonlinear kernel hyper-parameters. This is due to the challenges induced by optimizing the nonlinear kernel hyper-parameters. SVM with nonlinear kernels usually require either storing the kernel matrix in memory or computing it on the fly whenever it's needed. The kernel matrix has the same dimension as the number of training samples. Therefore, storing the kernel matrix for large datasets has a high memory requirement while computing it as needed is computationally

expensive. When we formulate the problem of tuning kernel hyper-parameters as an optimization problem, it often leads to a hard optimization problem with a highly nonlinear objective function and constraints. Further, performing the kernel estimation while optimizing the kernel hyper-parameter is computationally expensive, especially for large datasets.

Being able to use nonlinear kernels, is one of the main reasons that SVM is a powerful and popular machine learning method, and choosing the right value for the kernel hyper-parameters is crucial to obtain SVM models that work well. Therefore, being able to tune the kernel hyper-parameters together with the regularization parameters automatically is challenging but extremely useful. To optimize the kernel hyper-parameters, we can exploit the special structure of the kernel functions and apply gradient-based methods in a bilevel optimization framework. To deal with the nonlinearity in the optimization problem, we can remove the nonlinearity by approximation for the nonlinear terms, for instance, using a piece-wise linear function. Approximated gradient estimates can then be estimated instead again using the implicit differentiation trick. We are currently investigating the exact optimization problem formulation for optimizing the kernel hyper-parameters using the above approach.

Appendix

Data processing is often necessary and important before performing the analysis. The author has also spent great effort on data processing for the two projects in Chapter 2-3. Next, I will describe the specific data processing techniques I employed for each work.

I also included other supplemental material in this chapter, such as how to use the prediction model for a particular patient in practice.

7.1 Heart Failure Patient Readmission

7.1.1 Data Cleaning and Feature Engineering

Many of the data cannot be used directly from the heart failure patient database. Features need to be derived from the raw data. Therefore, in the beginning, a significant amount of time and effort was spent on cleaning the data and deriving features from the raw data. I performed three types of data cleaning and feature creating

1. Obtaining numerical values: For example, I need to extract systolic blood pressure and diastolic blood pressure from blood pressure entry, convert weight and lab measures into the same unit.

2. Feature categorization: Feature categorization for categorical features is another data cleaning process. Some categorical variables have too many values, but most of those values have small frequencies in the dataset. We manually aggregated those categorical variables into fewer categories using expert knowledge.

3. Label generation: I derived a set readmission flags also a primary diagnosis label from the raw data. Those flags are our prediction target. They are:

    (a) Readmission within 30 days after discharge due to HF

    (b) Readmission within study window due to HF

    (c) All-cause readmission

    (d) All-cause readmission within 30 days after discharge

    (e) Readmission due to HF as the primary diagnosis

    (f) Readmission within 30 days due to HF as the primary diagnosis

    (g) The time duration between admission and readmission

(h) Whether heart failure is an encounter's primary diagnosis

4. Feature engineering for complex features: For a complex feature like lab measures and procedures that have multiple repeated numerical values or a set of categorical values, I computed the frequency of measurement per day as its feature, e.g., the number of blood draws per day. For time series data including lab measures and vital signs, I created a feature template that computes detailed summary statistics described in Chapter 2 including gradients, mean, spread, maximal, and minimal value of the time series.

5. Comorbidity features: I created a class that can take a list of ICD-9 codes and return an integer Charlson index based an ICD-9 code and Charlson scoring dictionary. Finally, I created Charlson index for three features: diagnoses, diagnosis history, problem list.

6. Feature encoding: Categorical features do not have numerical order. The common way to deal with categorical data is to create dummy variables for each unique value of the categorical features. Then the dummy variable consists of binary values, which is the procedure I followed for feature encoding.

7.1.2 Missing Data Imputation

Another challenge is how to deal with missing values in our data set. We have missing values in both categorical features and continuous features.

First, I looked at the percentage of missing values for each feature. If the percentage of missing values for a certain feature exceeds 20%, I simply removed that feature instead of imputing the missing values. If the missing value is less than 20% for a certain feature, I imputed the missing value.

Categorical features don't have a significant amount of missing values and most categorical data are socio-demographical information. There are two obvious ways to treat missing values for categorical features. The first approach is that we simply treat missing values of that categorical feature as another dummy variable, which we may call it unknown. The second approach is similar to the imputing method we applied to continuous features. We can impute using the mode of the categorical feature or using its k nearest neighbors. For this dissertation, I treated missing values as a new dummy variable.

However, clinical data such as lab measurements and procedures have a significant amount of missing values most likely because that patient simply doesn't have particular lab measurement or procedure. For features representing a count of certain clinical measures, a missing value can be treated easily by assigning a 0 to it. However, for features representing values of actual clinical measures such as the value of Troponin T or value of blood pressure, missing values need to be imputed. The way we impute these missing values is simply imputing by mean value across all the encounters. The disadvantage of this approach is that it decreases the variances for that feature. Also, imputing by means is susceptible to outliers.

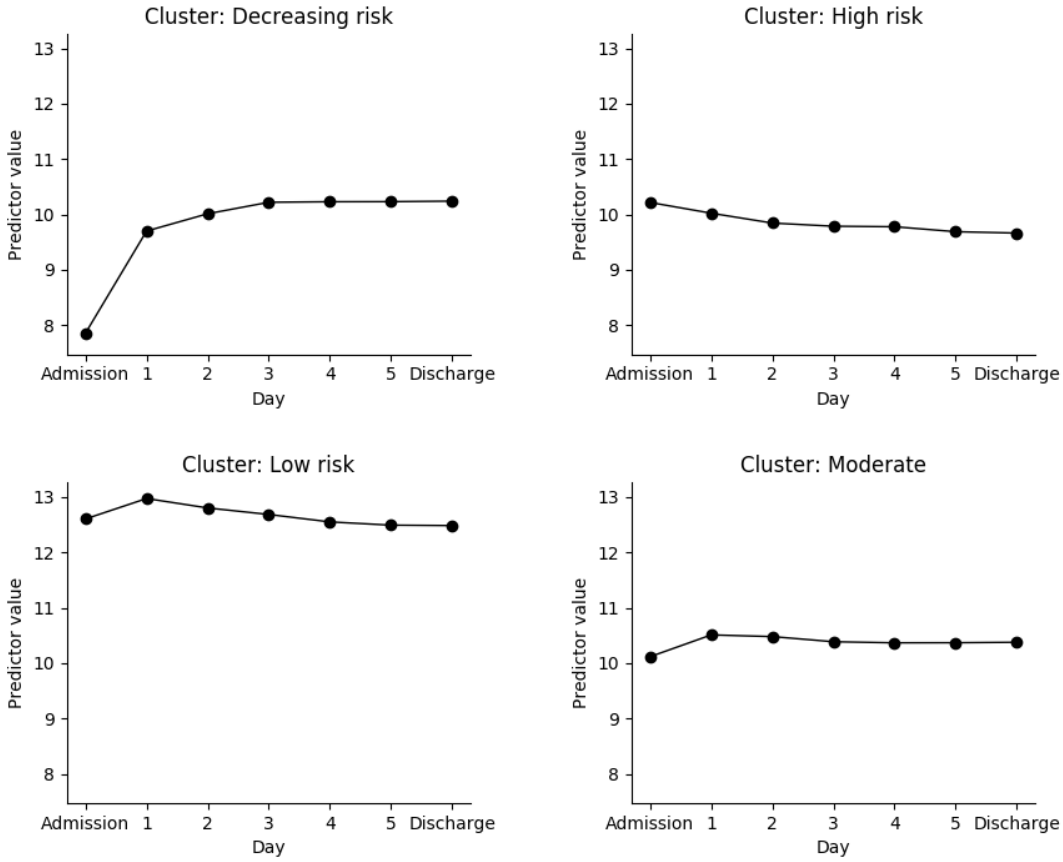
7.1.3 Data Transformation

We often need to performance data transformation to improve optimization efficiency and machine learning performance. I applied two types data transformation to our data set:

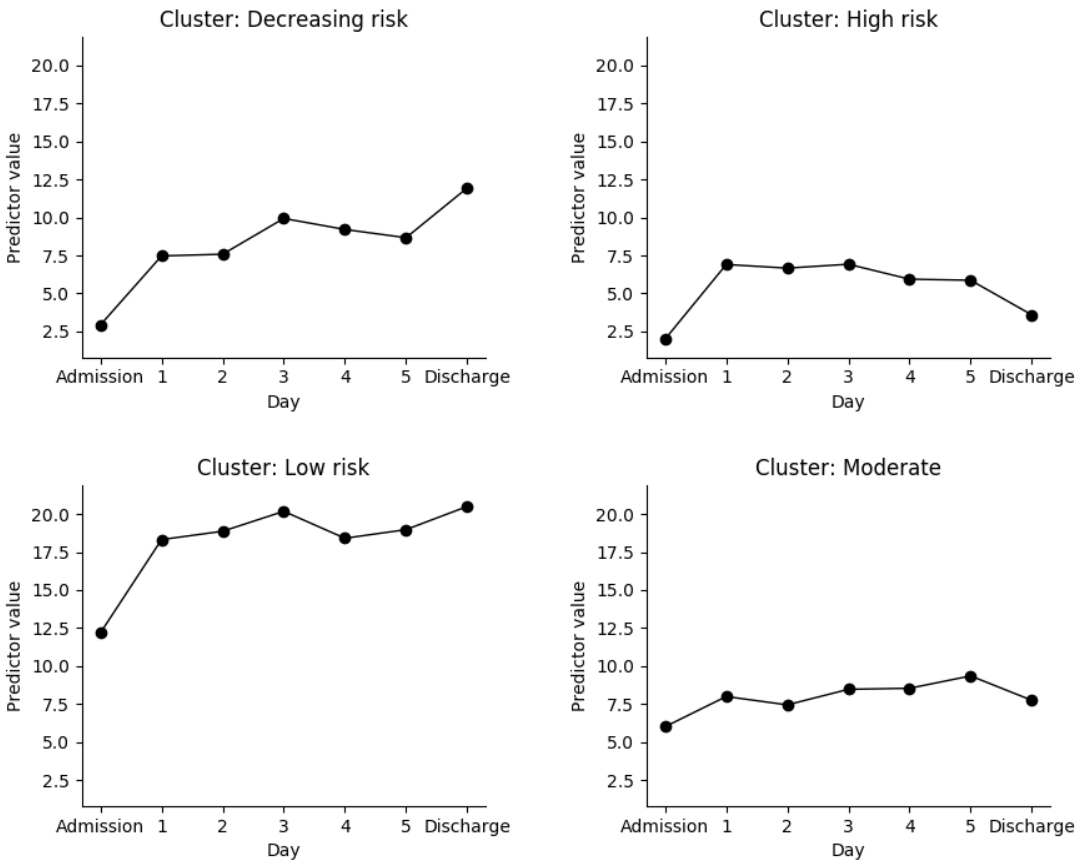
1. Feature scaling: The main preprocessing step for continuous features are feature scaling or feature standardization. Each features value may have very different scales and variance. To let the optimization procedure of the classifier, treat each feature equally, we need to standardize or scale our features. Gradient descent method will also converge faster with feature scaling. Otherwise, the features that have large values and variance will dominate the classification performance. The only classification methods that are feature scaling invariant are tree-based, such as Random Forest. Other classification methods like logistic regression and support vector machines need feature scaling. There are many different ways to scale or standardize features. The common ones are normalization and min-max scaling. Normalization normalizes the features to have mean 0 and unit variance. Minmax scaling scales the feature values to be within the range [0,1], a linear transformation. There is no obvious reason to choose one feature scaling method over the other. It depends on the application. In this dissertation, I use Z-score normalization. Unlike continuous features, categorical features don't need feature scaling.

2. Feature transformation: We observed that most of the continuous features (lab measures, vital signs, etc.) are not normally distributed, but rather appear to follow a log-normal distribution. Some machine learning algorithms depend on the assumption that the features are normally distributed such as Gaussian Naive Bayes. So to improve classification performance, I experimented with log transformation as a data preprocessing step by taking the natural logarithm of those continuous features.

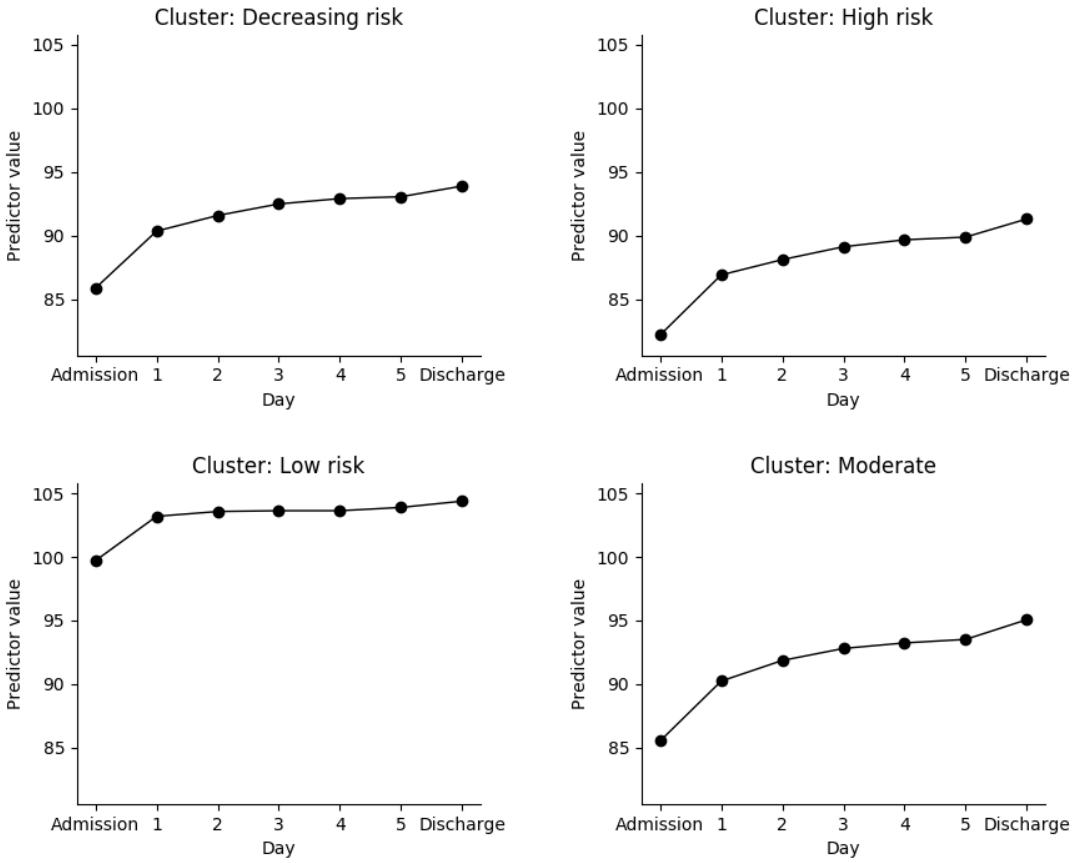
7.1.4 Trend of Dynamic Predictors



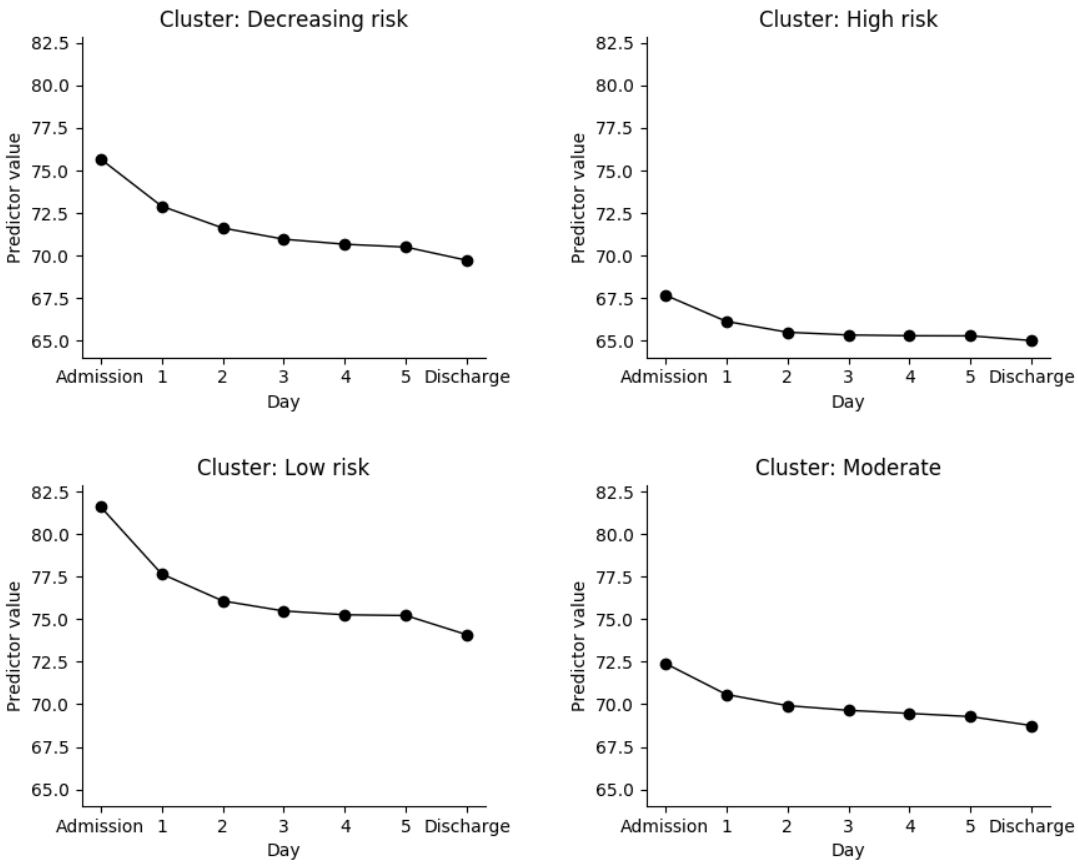
(a) Standard deviation of diastolic blood pressure (mmHg)



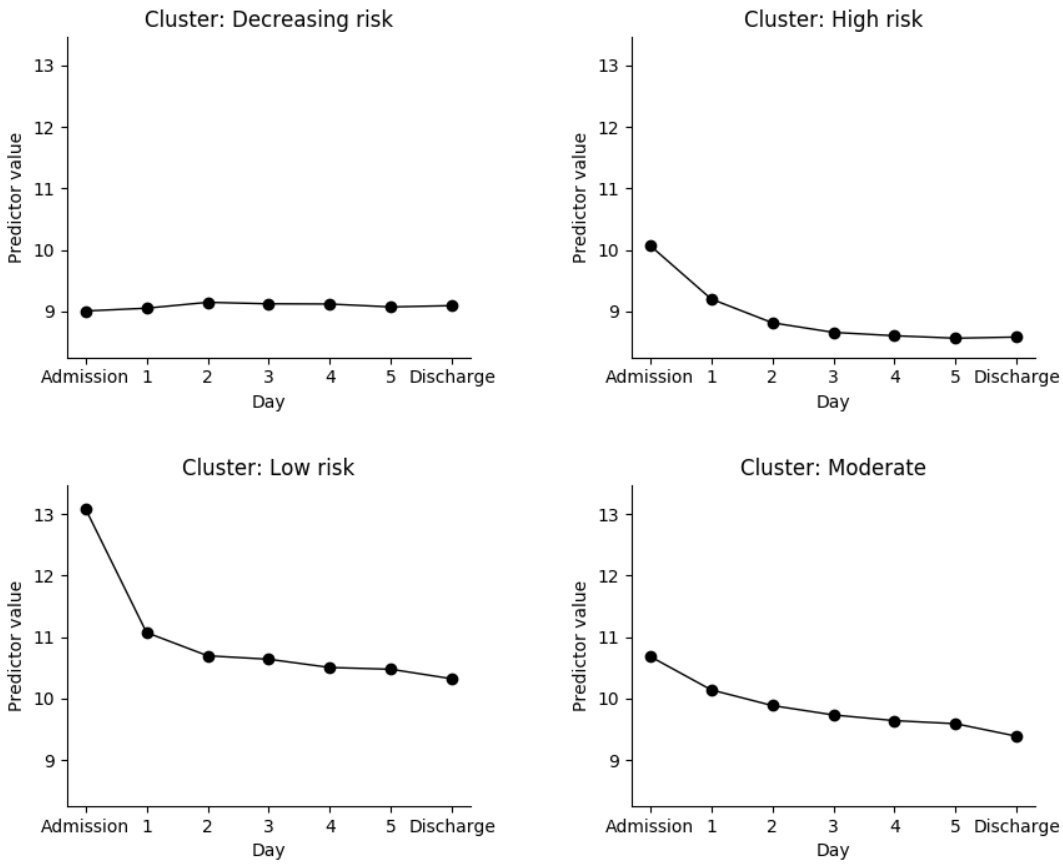
(b) Decrease of diastolic blood pressure level from admission (mmHg)



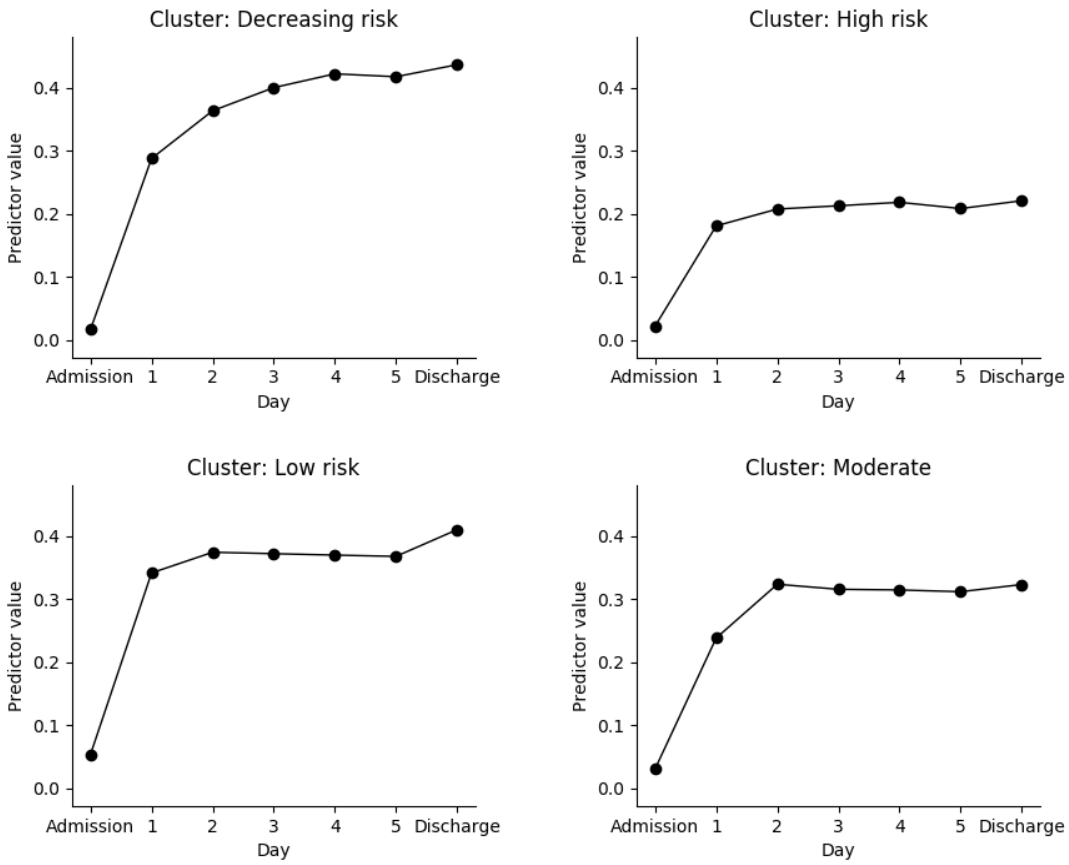
(c) Maximal diastolic blood pressure (mmHg)



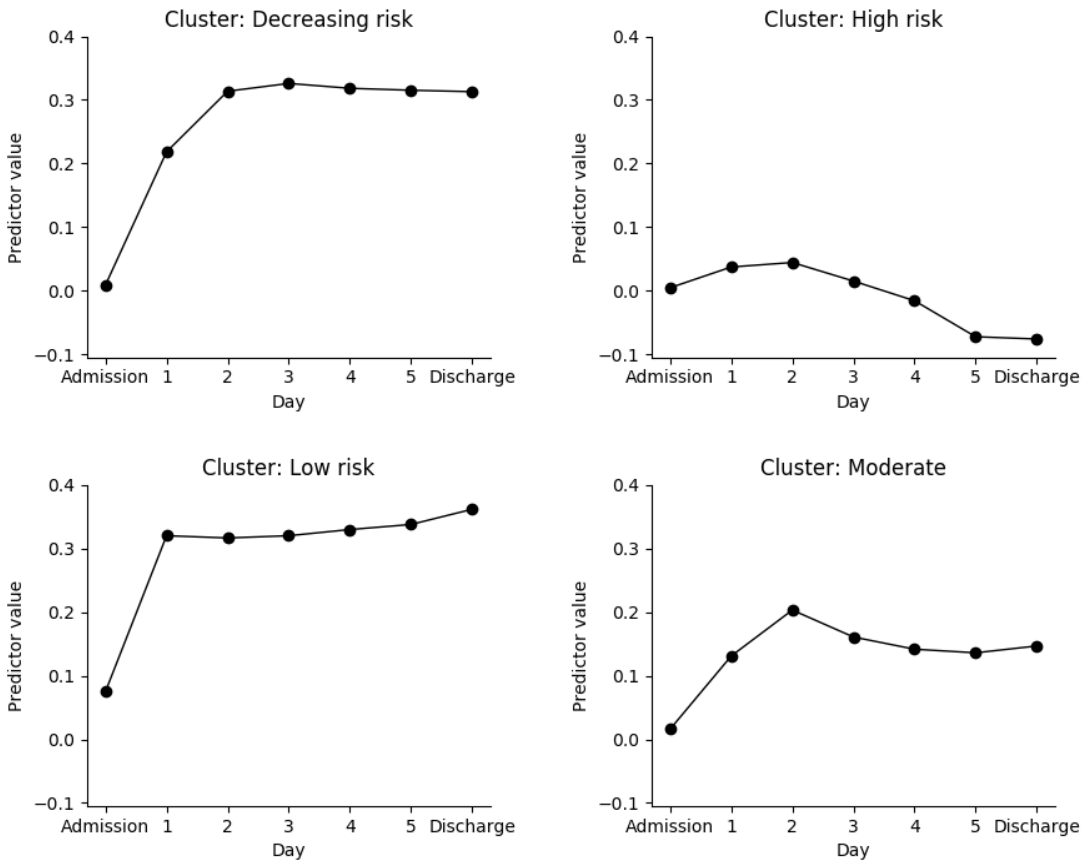
(d) Average diastolic blood pressure (mmHg)



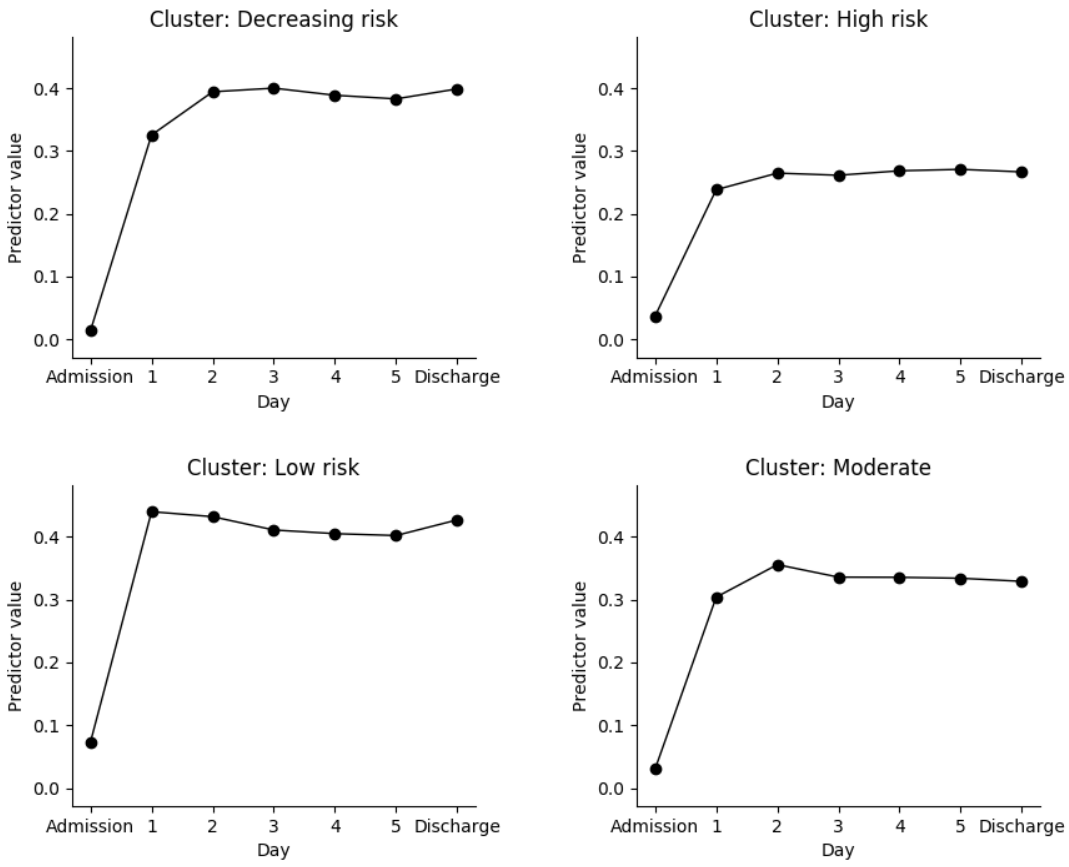
(e) Average absolute change of diastolic blood pressure (mmHg)



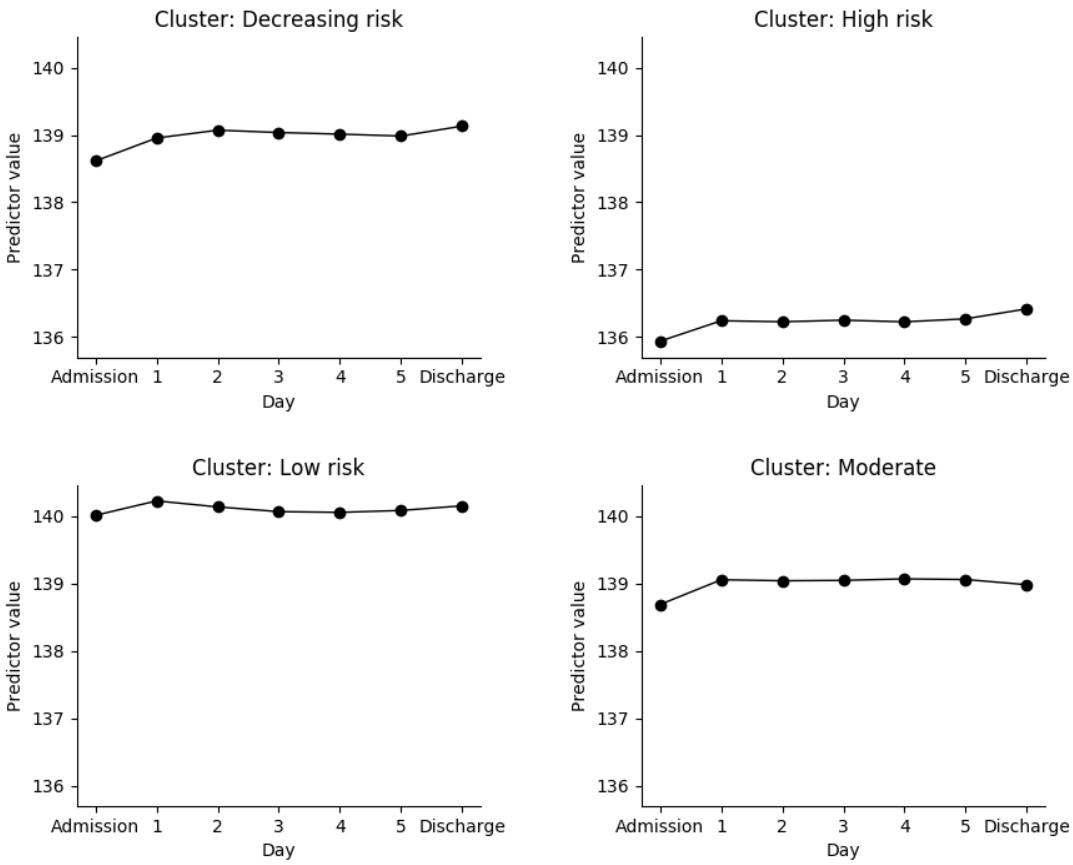
(f) Normalized time of minimal potassium starting from admission



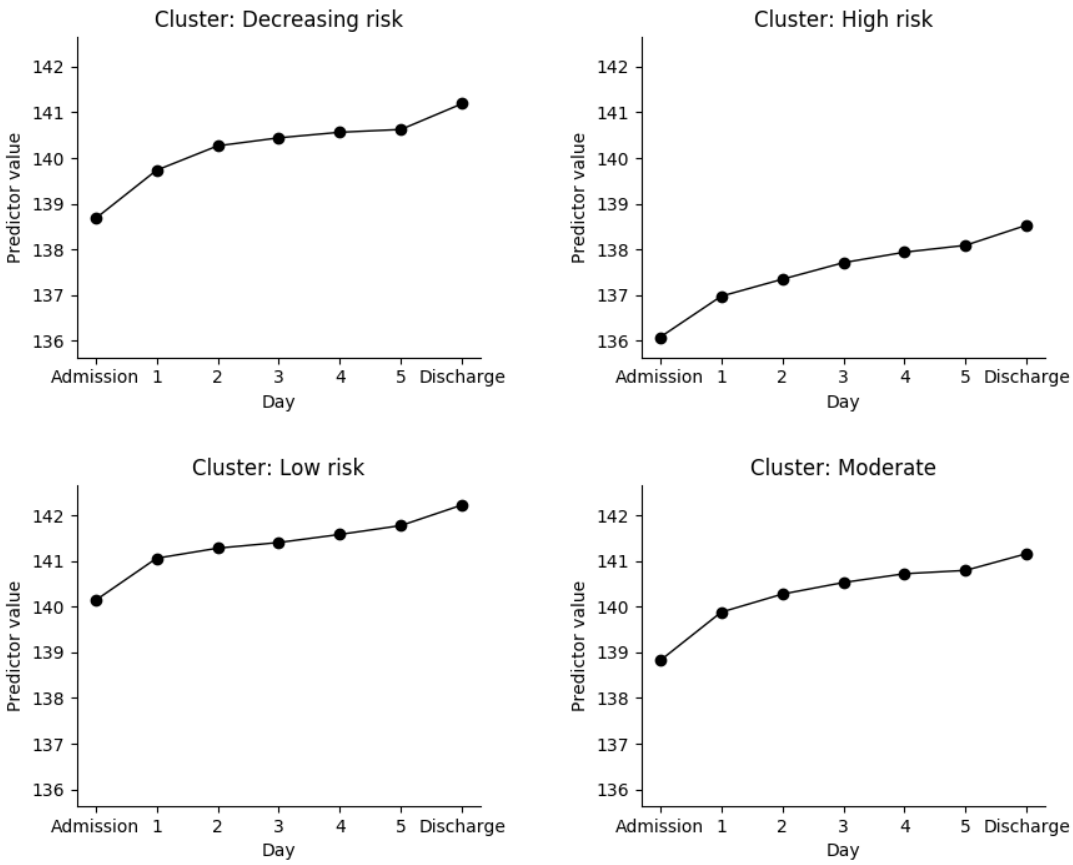
(g) Decrease of potassium level from admission (mmol/L)



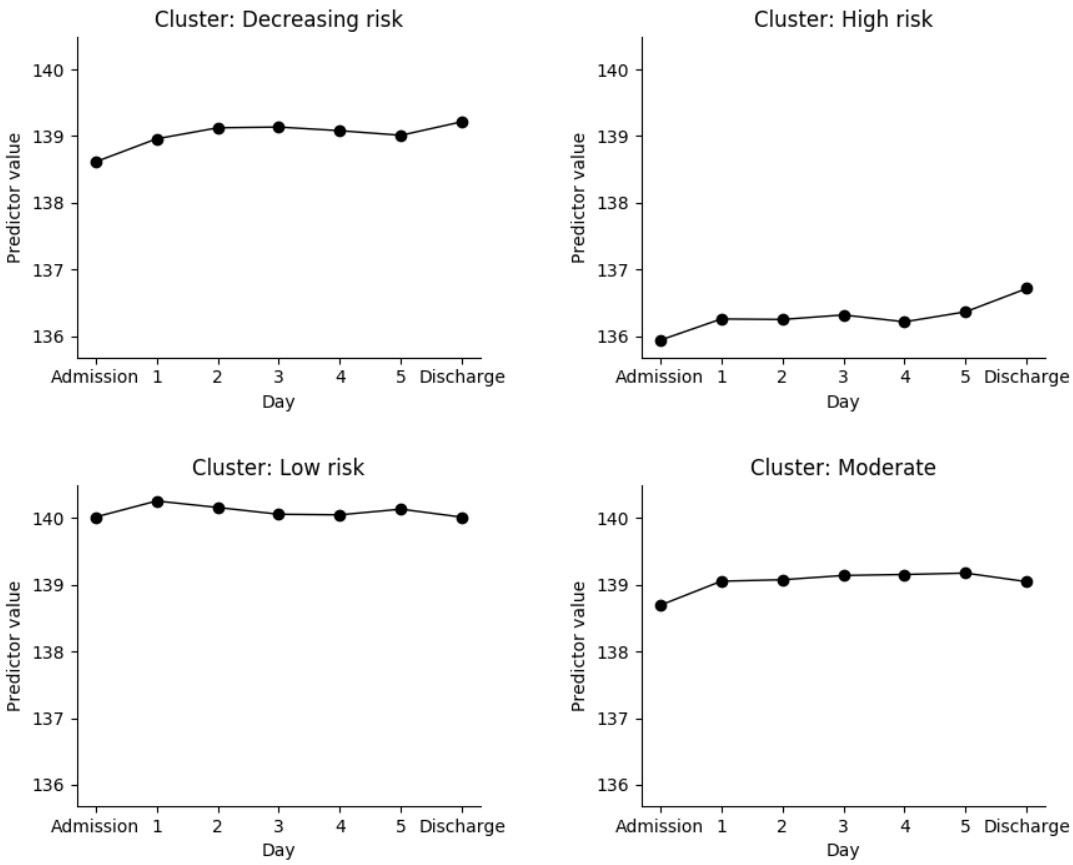
(h) Average absolute change of potassium (mmol/L)



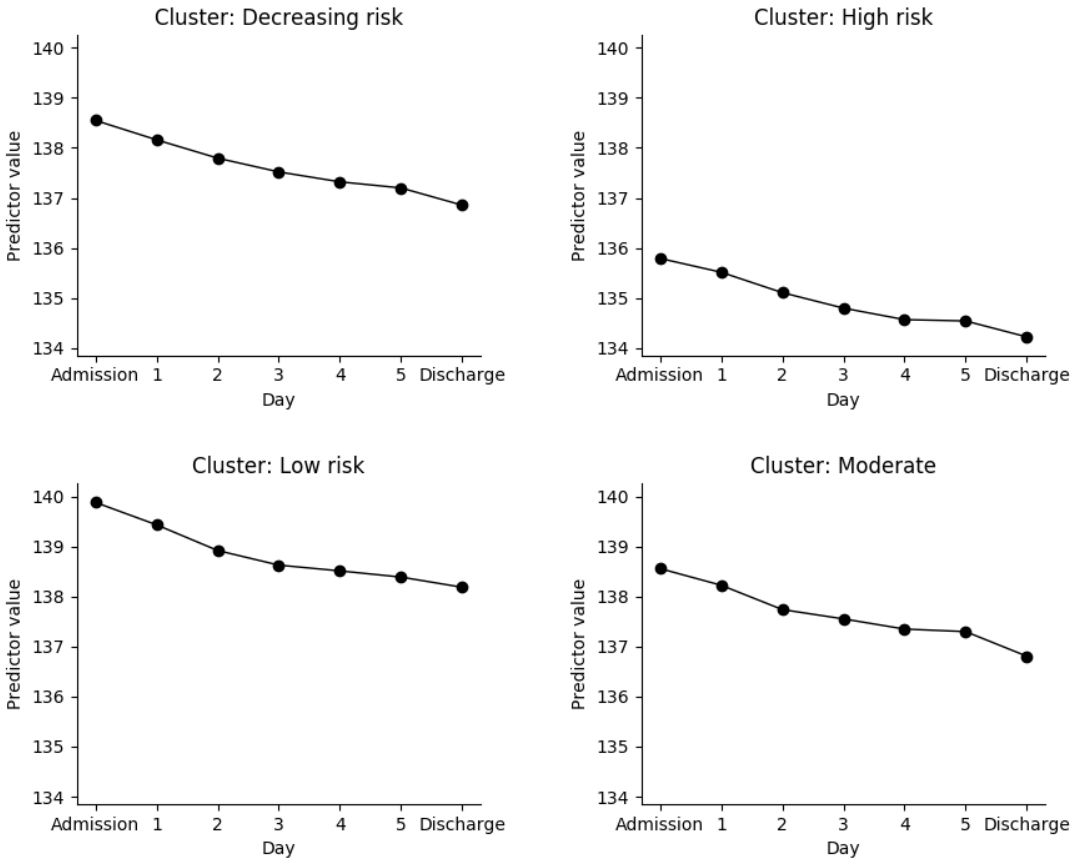
(i) Average sodium (mmol/L)



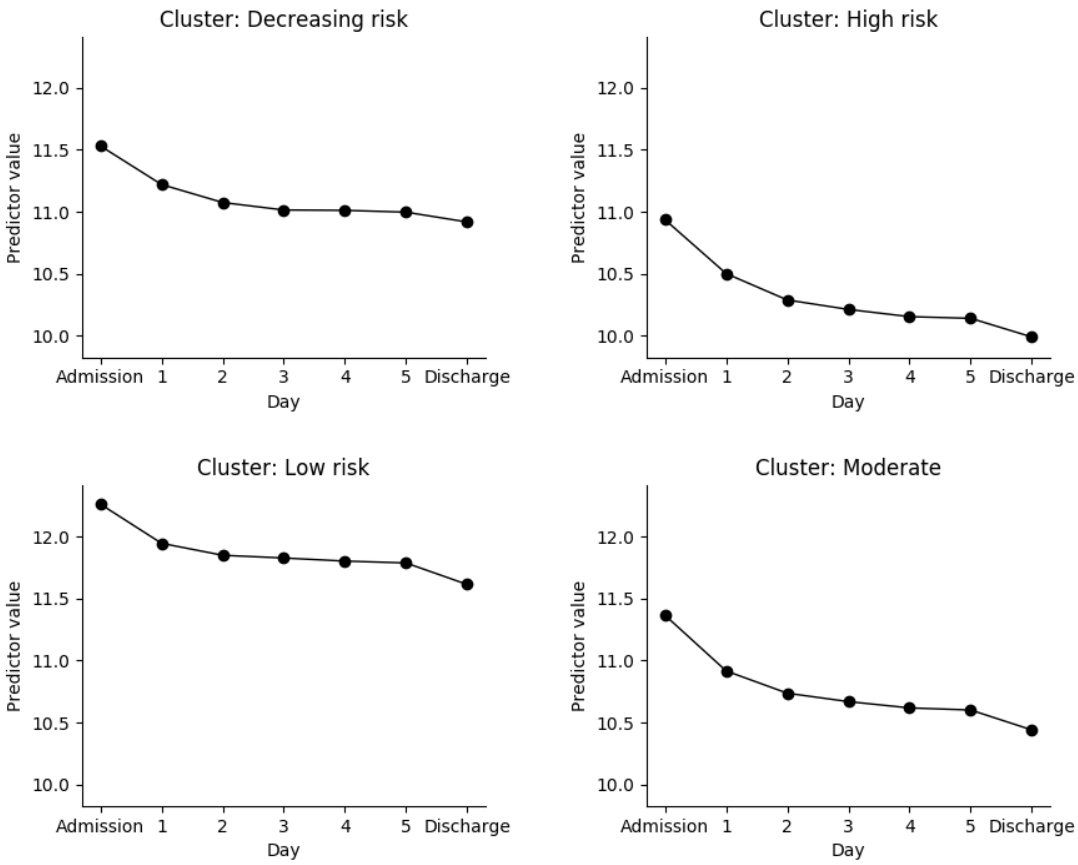
(j) Maximal sodium (mmol/L)



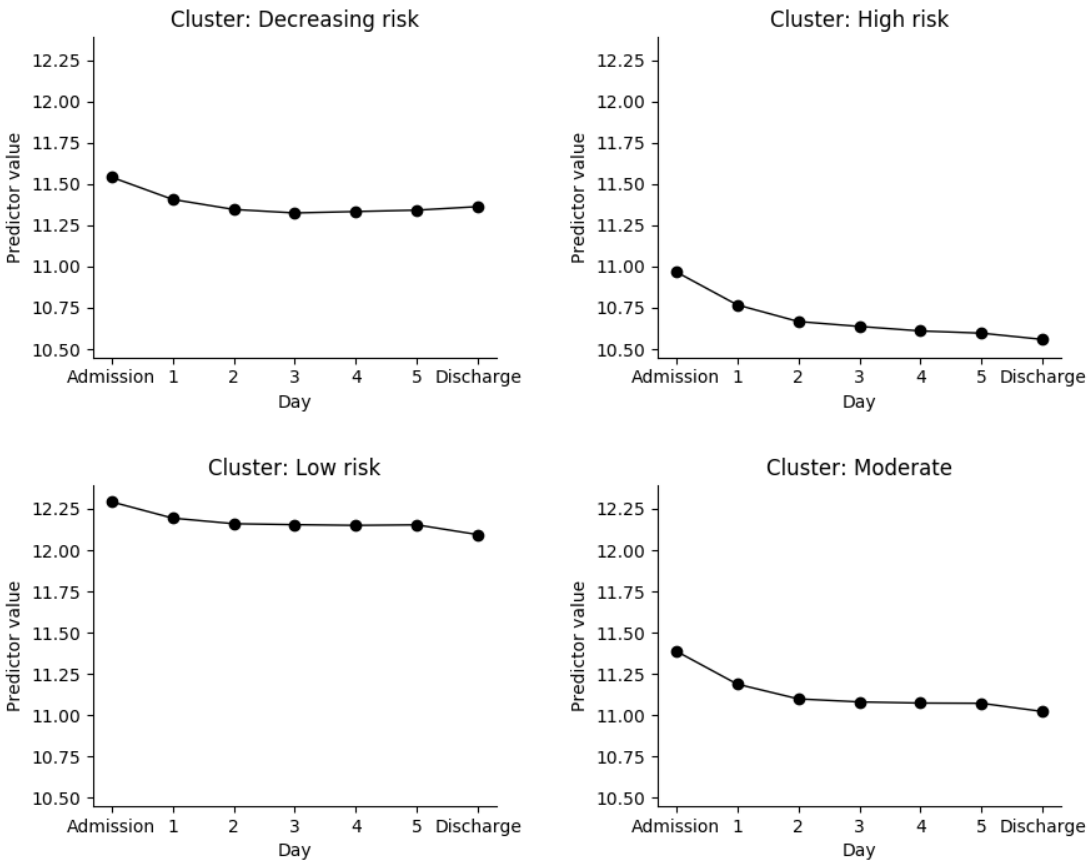
(k) Average value of last three sodium (mmol/L)



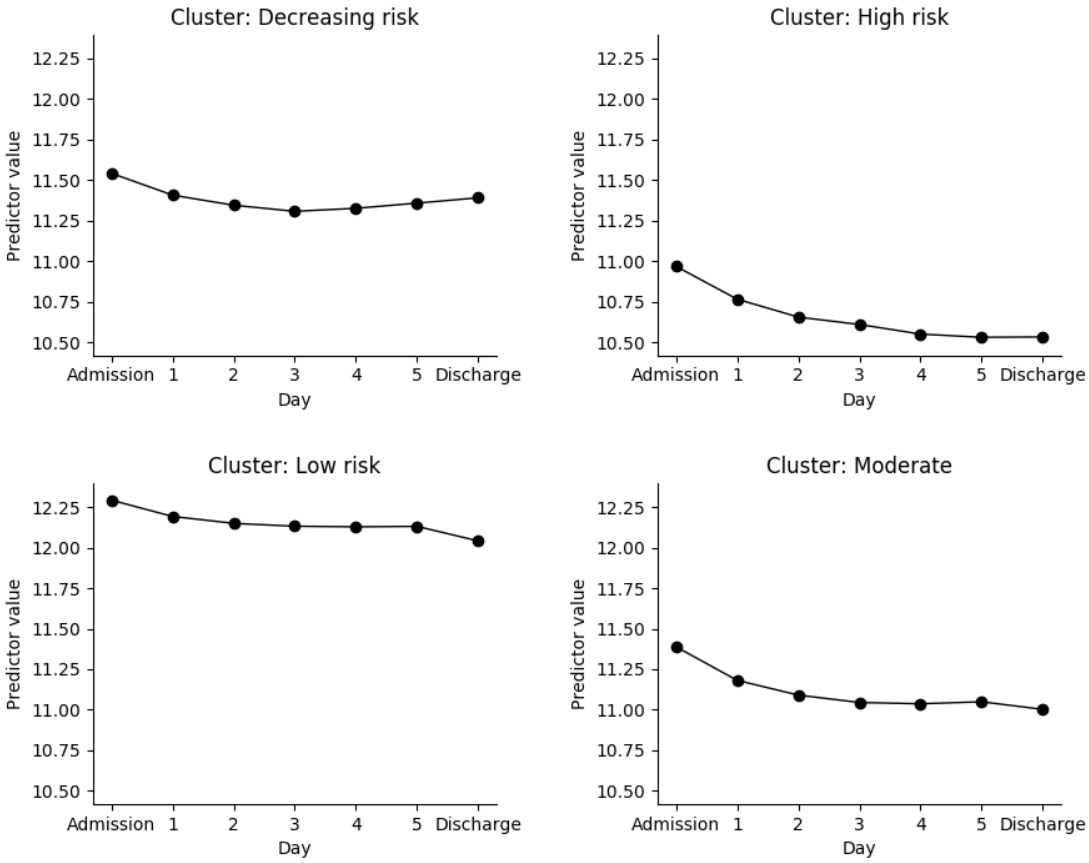
(l) Minimal sodium (mmol/L)



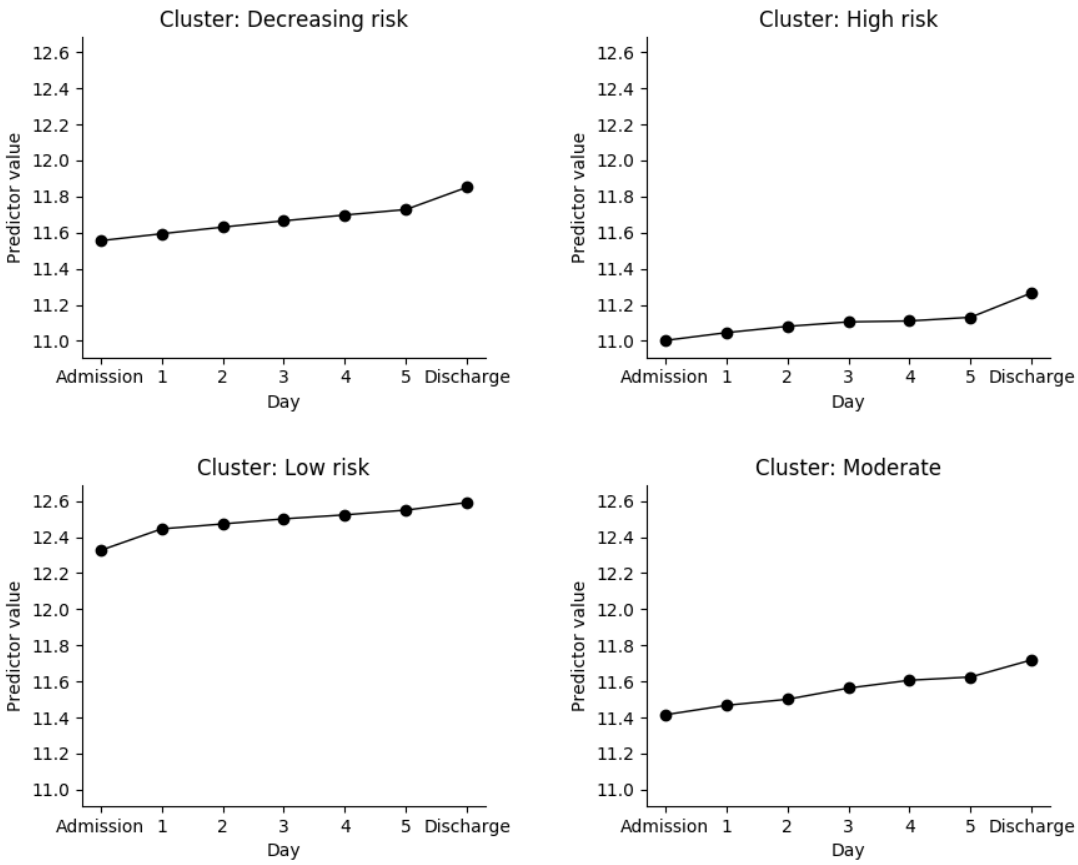
(m) Minimal hemoglobin (gm/dL)



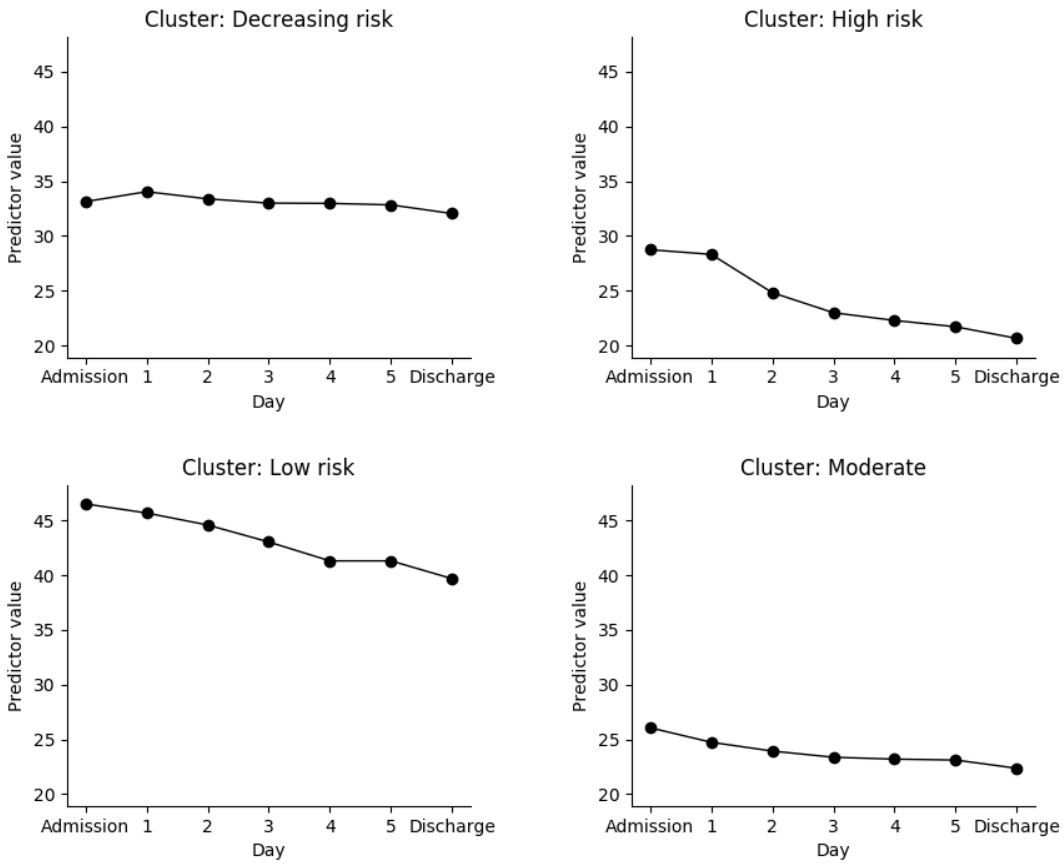
(n) Average hemoglobin (gm/dL)



(o) Average value of the last three hemoglobin (gm/dL)



(p) Maximal hemoglobin (gm/dL)



(q) Minimal ALT (Units/L)

Figure 7.1 Value of 17 dynamic discriminative predictors from admission to discharge.

7.1.5 Model Prediction for a Particular Patient

Using the model to predict the readmission risk for a particular patient is standard. After we input all the values of the predictors from that particular patient into the trained prediction model, we

can obtain the predicted readmission risk. However, in practice, we may still have follow-up tasks to perform after the prediction. For instance, the physician could want to know which features may be driving the high readmission risk for that patient if the predicted readmission risk is high. Knowing the important features for that patient is important to help the physician provide interventions to improve the patients' health condition and lower the readmission risk.

The specific steps for detecting the potential driving features for high readmission risk and providing interventions are as follows:

1. Go through all the dynamic lab tests and vital sign predictors in the model for that patient and find the ones that are out of the normal range. For instance, diastolic blood pressure higher than 90 mmHg indicates high blood pressure and we can mark this predictor for this patient. Let's denote this subset of abnormal features as $\boldsymbol{x}_a$.
2. For each predictor x_i in $\boldsymbol{x}_a$, we set the value of x_i to a normal value (for instance, 70 mmHg for diastolic blood pressure). Then, we rerun the prediction model while fixing the values for all the other predictors. If the predicted readmission risk decreases, we mark this predictor as a potential predictor that can lead to interventions. We denote the set of these marked predictors as $\boldsymbol{x}_m$.
3. For the predictors $\boldsymbol{x}_m$, we can rank them based on the amount of reduced readmission risk caused by setting the value of each predictors to a normal value.

For instance, among all the predictors in $\boldsymbol{x}_m$, if setting the diastolic blood pressure to 70 mmHg leads to the lowest readmission risk, we rank diastolic blood pressure as the most important predictor for that patient.

4. The physicians can prioritize interventions based on the ranked list of important predictors $\boldsymbol{x}_m$ for that patient. For instance, first, provide interventions to bring down the patient's diastolic blood pressure level; second, provide interventions to increase the patient's hemoglobin level.

In the case where the physician is not able to provide interventions to bring back the value of important predictors to be exactly in the normal range, the odds ratio of each predictor can be used to rank those predictors in $\boldsymbol{x}_m$. A larger odds ratio indicates a more important predictor, and thus can be guide for intervention.

Although this provides a guideline for identifying important features for a particular patient, expert knowledge of physicians is needed to finally deliver interventions.

7.2 Radiation Oncology

7.2.1 Data Processing

For the radiation oncology work, two main data processing techniques were studied and applied: outlier detection and missing data imputation. Outlier detection was performed on patients' weight outcome data. As xerostomia outcome was measured using a one to four scale, we didn't observe any outliers in xerostomia outcome data. However, we have observed obvious outliers in the weight data. The values of weight are continuous and can have large variations. Therefore, it's more prone to have outliers. A missing data imputation technique was also conducted for longitudinal weight data but not for xerostomia. The concern is that we won't be able to reliably evaluate prediction model performance if the xerostomia outcome is imputed. Instead, we used last-observation-carry-forward to obtain the xerostomia outcomes for dropped-out patients. The popular method of multiple imputations by chained equations (MICE) was also explored for imputing missing features [144]. However, we didn't apply this method due to its theoretical inconsistencies with our analysis.

In the following section, I describe these two data processing techniques applied to the longitudinal weight outcome data for our patient cohort.

7.2.1.1 Outlier Detection

We used a metric called median absolute deviance (MAD) to detect the outliers. Assuming we have a time series of weight measurements for a patient, x_i, denote the median of x_i as $M(x_i)$. Then we have MAD $= median(|x_i - M(x_i)|)$. Let's define a threshold parameter γ to detect the outliers. If the deviance of one weight measurement from its median as a portion of the MAD is

large than γ, i.e., $\frac{|x_i - M(x_i)|}{MAD} \geq \gamma$, we detected it as an outlier. Figure 7.2 shows the results of outlier detection using this method. The left figure is the raw data that contains outliers (shown as round dots), and the right figure shows the weight time series data after removing the outliers.

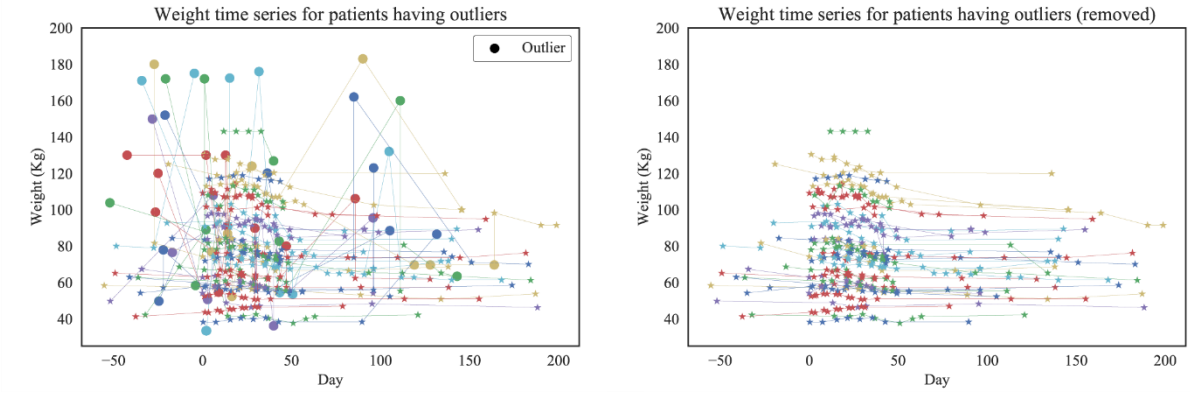


Figure 7.2 Outlier detection for longitudinal weight outcomes.

7.2.1.2 Missing Data Imputation

For imputing missing longitudinal weight data, we applied two strategies. If the data is missing in the middle of two available measurements, we used linear interpolation to impute the missing data. If the data is missing due to dropouts, we imputed it using a dependent censoring model by Andrea and Robins (1995) [145]. Figure 7.3 and Figure 7.4 presents the results of applying this imputation method on the weight data. It shows the distribution of imputed weights is slightly smaller than

the complete case. This illustrates why using a complete case will lead to a small bias in the sample, and this method can be used to correct this bias.

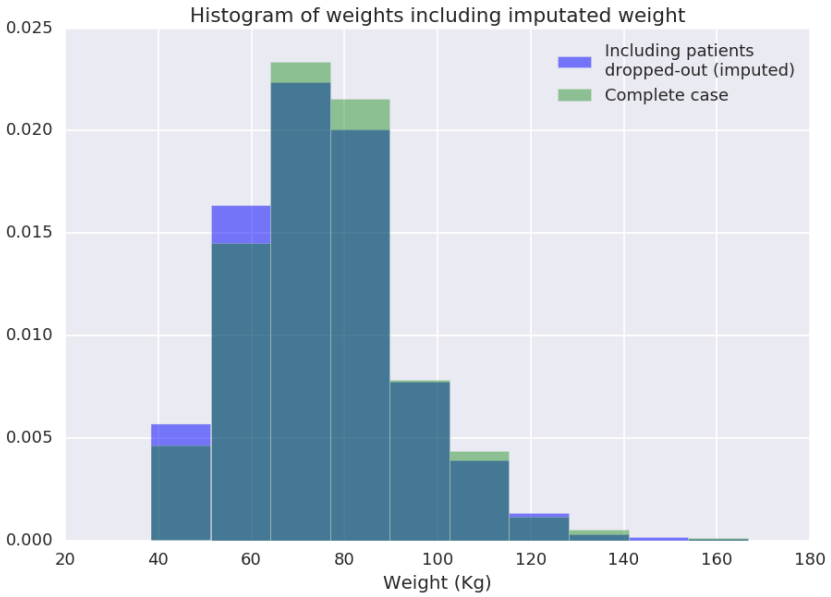


Figure 7.3 Histograms comparing the weight distribution between complete cases and cases including imputed weights. Weights were imputed for patients who dropped out.

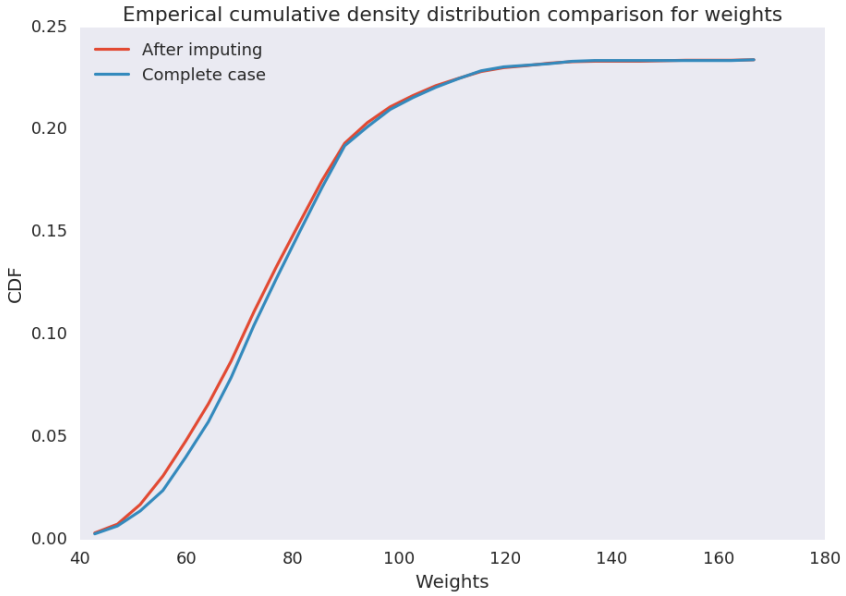


Figure 7.4 Cumulative density distributions comparing the weight distribution between complete cases and cases including imputed weights. Weights were imputed for dropped-out patients.

Bibliography

1 Russell S, Norvig P. *Artificial intelligence : a modern approach*. Prentice Hall 2010.

2 Friedman J, Hastie T, Tibshirani R. *The elements of statistical learning*. New York: Springer series in statistics 2001.

3 Cortes C, Vapnik V. Support-vector networks. *Mach Learn* 1995;**20**:273–97. doi:10.1007/BF00994018

4 Nocedal J, Wright SJ. *Numerical optimization*. Springer 2006.

5 Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982;**143**:29–36. doi:10.1148/radiology.143.1.7063747

6 Hastie T, Tibshirani R, Wainwright M. *Statistical learning with Sparsity : the lasso and generalizations*.

7 Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw* 2010;**33**:1–20. doi:10.18637/jss.v033.i01

8 Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the

gap statistic. *J R Stat Soc Ser B (Statistical Methodol* 2001;**63**:411–23. doi:10.1111/1467-9868.00293

9 Bottou L. Large-Scale Machine Learning with Stochastic Gradient Descent. In: *Proceedings of COMPSTAT'2010*. Heidelberg: : Physica-Verlag HD 2010. 177–86. doi:10.1007/978-3-7908-2604-3_16

10 Rumelhart DE, McClelland JL, University of California SDPRG. *Parallel distributed processing : explorations in the microstructure of cognition*. MIT Press 1986. https://dl.acm.org/citation.cfm?id=104293 (accessed 6 Mar 2018).

11 Wright SJ, J. S. Coordinate descent algorithms. *Math Program* 2015;**151**:3–34. doi:10.1007/s10107-015-0892-3

12 Hsieh C-J, Chang K-W, Lin C-J, *et al.* A Dual Coordinate Descent Method for Large-scale Linear SVM. http://delivery.acm.org/10.1145/1400000/1390208/p408-hsieh.pdf?ip=162.129.251.103&id=1390208&acc=ACTIVE SERVICE&key=7777116298C9657D.34B115928DB6308C.4D4702B0C3E38B35.4D4702B0C3E38B35&CFID=1016046698&CFTOKEN=65777655&__acm__=1515171928_5ab1ba57334103616b7 (accessed 5 Jan 2018).

13 Chioncel O, Greene SJ, Vaduganathan M. The Global Health and Economic Burden of Hospitalizations for Heart Failure Lessons Learned From Hospitalized Heart Failure Registries. *J Am Coll Cardiol* 2014;**63**:1123–33. doi:10.1016/j.jacc.2013.11.053

14 Centers for Medicare and Medicaid Services. Readmissions Reduction Program (HRRP). https://www.cms.gov/medicare/medicare-fee-for-service-payment/acuteinpatientpps/readmissions-reduction-program.html (accessed 1 Jan 2017).

15 Futoma J, Morris J, Lucas J. A comparison of models for predicting early hospital readmissions. *J Biomed Inform* 2015;**56**:229–38. doi:10.1016/j.jbi.2015.05.016

16 Cholleti S, Post A, Gao J, *et al.* Leveraging derived data elements in data analytic models for understanding and predicting hospital readmissions. *AMIA Annu Symp Proc* 2012;**2012**:103–11.http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3540449&tool=pmcentrez&rendertype=abstract

17 Lemke KW, Weiner JP, Clark JM. Development and validation of a model for predicting inpatient hospitalization; 22002640. *Med Care* 2012;**50**:131–9. doi:10.1097/MLR.0b013e3182353ceb

18 Hammill BG, Curtis LH, Fonarow GC, *et al.* Incremental value of clinical data beyond claims data in predicting 30-day outcomes after heart failure hospitalization. *Circ Cardiovasc Qual Outcomes* 2011;**4**:60–7. doi:10.1161/CIRCOUTCOMES.110.954693

19 Meadem N, Verbiest N, Zolfaghar K, *et al.* Exploring preprocessing techniques for prediction of risk of readmission for congestive heart failure patients. *Data Min ...* Published Online First: 2013.http://chbrown.github.io/kdd-2013-

usb/workshops/DMH/doc/dmh2145_Meadem.pdf

20 Tang Y-D, Katz SDM. Anemia in Chronic Heart Failure: Prevalence, Etiology, Clinical Correlates, and Treatment Options. *Circulation* 2006;**113**:2454–61. doi:10.1161/CIRCULATIONAHA.105.583666

21 Saczynski JS, Andrade SE, Harrold LR, *et al.* A systematic review of validated methods for identifying heart failure using administrative data. *Pharmacoepidemiol Drug Saf* 2012;**21**:129–40. doi:10.1002/pds.2313

22 Huynh QL, Saito M, Blizzard CL, *et al.* Roles of nonclinical and clinical data in prediction of 30-day rehospitalization or death among heart failure patients. *J Card Fail* 2015;**21**:374–81. doi:10.1016/j.cardfail.2015.02.002

23 Flint KM, Allen LA, Pham M, *et al.* B-type natriuretic peptide predicts 30-day readmission for heart failure but not readmission for other causes. *J Am Heart Assoc* 2014;**3**:1–10. doi:10.1161/JAHA.114.000806

24 Januzzi JL, Van Kimmenade R, Lainchbury J, *et al.* NT-proBNP testing for diagnosis and short-term prognosis in acute destabilized heart failure: An international pooled analysis of 1256 patients: The international collaborative of NT-proBNP study. *Eur Heart J* 2006;**27**:330–7. doi:10.1093/eurheartj/ehi631

25 Amarasingham R, Moore BJ, Tabak YP, *et al.* An automated model to identify heart failure patients at risk for 30-day readmission or death using electronic medical record data.

2010;**48**. doi:10.1097/MLR.0b013e3181ef60d9

26 Nguyen OK, Makam AN, Clark C, *et al.* Vital Signs Are Still Vital : Instability on Discharge and the Risk of Post-Discharge Adverse Outcomes. *J Gen Intern Med* doi:10.1007/s11606-016-3826-8

27 Kansagara D, Englander H, Salanitro A, *et al.* Risk Prediction Models for Hospital Readmission: A systematic Review. 2013;**306**:1688–98. doi:10.1001/jama.2011.1515.Risk

28 Dugas AF, Kirsch TD, Toerper M, *et al.* An Electronic Emergency Triage System to Improve Patient Distribution by Critical Outcomes. *J Emerg Med* 2016;**50**:910–8. doi:10.1016/J.JEMERMED.2016.02.026

29 Levin S, Toerper M, Hamrock E, *et al.* Machine-Learning-Based Electronic Triage More Accurately Differentiates Patients With Respect to Clinical Outcomes Compared With the Emergency Severity Index. *Ann Emerg Med* Published Online First: 6 September 2017. doi:10.1016/J.ANNEMERGMED.2017.08.005

30 Wiens J, Guttag J V, Horvitz E. Patient Risk Stratification for Hospital-Associated C. diff as a Time-Series Classification Task. *Adv Neural Inf Process Syst 25 (NIPS 2012)* 2012;:1–9.

31 Friedman J, Hastie T, Tibshirani R. *The elements of statistical learning*. Springer series in statistics Springer, Berlin 2001.

32 Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat*

2001;**29**:1189–232.https://projecteuclid.org/download/pdf_1/euclid.aos/1013203451 (accessed 21 Dec 2017).

33 Cribari-Neto F, Zeileis A. Beta Regression in R. *J Stat Softw* 2010;**34**:1–24. doi:10.18637/jss.v034.i02

34 e-Handbook of statistical methods. Quantile-Quantile Plot. 2004.https://www.itl.nist.gov/div898/handbook/eda/section3/qqplot.htm (accessed 12 Apr 2018).

35 Gibbons JD, Chakraborti S. *Nonparametric Statistical Inference*. 4th ed. CRC Press 2003. http://erecursos.uacj.mx/bitstream/handle/20.500.11961/2064/Gibbons%2C 2003.pdf?sequence=14&isAllowed=y (accessed 12 Apr 2018).

36 Gutierrez C, Blanchard DG. Diastolic heart failure: Challenges of diagnosis and treatment. *Am Fam Physician* 2004;**69**:2609–16.

37 Al-Ahmad a, Rand WM, Manjunath G, *et al.* Reduced kidney function and anemia as risk factors for mortality in patients with left ventricular dysfunction. *J Am Coll Cardiol* 2001;**38**:955–62.

38 Anand IS, Kuskowski MA, Rector TS, *et al.* Anemia and change in hemoglobin over time related to mortality and morbidity in patients with chronic heart failure: Results from Val-HeFT. *Circulation* 2005;**112**:1121–7. doi:10.1161/CIRCULATIONAHA.104.512988

39 Androne AS, Katz SD, Lund L, *et al.* Hemodilution is common in patients with advanced

heart failure. *Circulation* 2003;**107**:226–9. doi:10.1161/01.CIR.0000052623.16194.80

40 Brucks S, Little WC, Chao T, *et al.* Relation of anemia to diastolic heart failure and the effect on outcome. *Am J Cardiol* 2004;**93**:1055–7. doi:10.1016/j.amjcard.2003.12.062

41 Ezekowitz JA, McAlister FA, Armstrong PW. Anemia is common in heart failure and is associated with poor outcomes: Insights from a cohort of 12 065 patients with new-onset heart failure. *Circulation* 2003;**107**:223–5. doi:10.1161/01.CIR.0000052622.51963.FC

42 Felker GM, Gattis WA, Leimberger JD, *et al.* Usefulness of anemia as a predictor of death and rehospitalization in patients with decompensated heart failure. *Am J Cardiol* 2003;**92**:625–8. doi:10.1016/S0002-9149(03)00740-9

43 Sharma R, Francis DP, Pitt B, *et al.* Haemoglobin predicts survival in patients with chronic heart failure: A substudy of the ELITE II trial. *Eur Heart J* 2004;**25**:1021–8. doi:10.1016/j.ehj.2004.04.023

44 Aronson D, Mittleman M a, Burger AJ. Elevated blood urea nitrogen level as a predictor of mortality in patients admitted for decompensated heart failure. *Am J Med* 2004;**116**:466–73. doi:10.1016/j.amjmed.2003.11.014

45 Damman K, Navis G, Voors AA, *et al.* Worsening Renal Function and Prognosis in Heart Failure: Systematic Review and Meta-Analysis. *J Card Fail* 2007;**13**:599–608. doi:10.1016/j.cardfail.2007.04.008

46 Smith GL, Lichtman JH, Bracken MB, *et al.* Renal Impairment and Outcomes in Heart

Failure. Systematic Review and Meta-Analysis. *J Am Coll Cardiol* 2006;**47**:1987–96. doi:10.1016/j.jacc.2005.11.084

47 Krumholz HM, Chen YT, Wang Y, *et al.* Predictors of readmission among elderly survivors of admission with heart failure. *Am Heart J* 2000;**139**:72–7. doi:10.1016/S0002-8703(00)90311-9

48 Donze, Jacques; Aujesky, Drahomir; Williams D, Schnipper JL. Potentially Avoidable 30-Day Hospital Readmissions in Medical Patients. *JAMA INTERN MED* 2013;**173**:632–8. doi:10.1001/jamainternmed.2013.3023

49 Hasan O, Meltzer DO, Shaykevich SA, *et al.* Hospital readmission in general medicine patients: A prediction model. *J Gen Intern Med* 2010;**25**:211–9. doi:10.1007/s11606-009-1196-1

50 Billings J, Dixon J, Mijanovich T, *et al.* Case finding for patients at risk of readmission to hospital: development of algorithm to identify high risk patients. *BMJ* 2006;**333**.http://www.bmj.com/content/333/7563/327 (accessed 2 Sep 2017).

51 Eisbruch A, Kim HM, Terrell JE, *et al.* Xerostomia and its predictors following parotid-sparing irradiation of head-and-neck cancer. *Int J Radiat Oncol Biol Phys* 2001;**50**:695–704. doi:10.1016/S0360-3016(01)01512-7

52 Dirix P, Nuyts S, Van Den Bogaert W. Radiation-induced xerostomia in patients with head and neck cancer: A literature review. *Cancer* 2006;**107**:2525–34. doi:10.1002/cncr.22302

53 Lee T-F, Liou M-H, Huang Y-J, *et al.* LASSO NTCP predictors for the incidence of xerostomia in patients with head and neck squamous cell carcinoma and nasopharyngeal carcinoma. *Sci Rep* 2015;**4**:6217. doi:10.1038/srep06217

54 Beetz I, Schilstra C, van der Schaaf A, *et al.* NTCP models for patient-rated xerostomia and sticky saliva after treatment with intensity modulated radiotherapy for head and neck cancer: The role of dosimetric and clinical factors. *Radiother Oncol* 2012;**105**:101–6. doi:10.1016/J.RADONC.2012.03.004

55 Deasy JO, Moiseenko V, Marks L, *et al.* Radiotherapy Dose-Volume Effects on Salivary Gland Function. *Int J Radiat Oncol Biol Phys* 2010;**76**:58–63. doi:10.1016/j.ijrobp.2009.06.090

56 Konings AWT, Faber H, Cotteleer F, *et al.* Secondary radiation damage as the main cause for unexpected volume effects: A histopathologic study of the parotid gland. *Int J Radiat Oncol Biol Phys* 2006;**64**:98–105. doi:10.1016/j.ijrobp.2005.06.042

57 Van Luijk P, Pringle S, Deasy JO, *et al.* Sparing the region of the salivary gland containing stem cells preserves saliva production after radiotherapy for head and neck cancer. *Sci Transl Med* 2015;**7**:305ra147. doi:10.1126/scitranslmed.aac4441

58 Nanduri LSY, Maimets M, Pringle SA, *et al.* Regeneration of irradiated salivary glands with stem cell marker expressing cells. *Radiother Oncol* 2011;**99**:367–72. doi:10.1016/j.radonc.2011.05.085

59 Robertson SP, Quon H, Kiess AP, *et al.* A data-mining framework for large scale analysis of dose-outcome relationships in a database of irradiated head and neck cancer patients. *Med Phys* 2015;**42**:4329–37. doi:10.1118/1.4922686

60 Nakatsugawa M, Cheng Z, Goatman KA, *et al.* Radiomic Analysis of Salivary Glands and Its Role for Predicting Xerostomia in Irradiated Head and Neck Cancer Patients. *Int J Radiat Oncol* 2016;**96**:S217. doi:10.1016/j.ijrobp.2016.06.539

61 Xuan Hui, Harry Quon, MS, Scott P Robertson, Zhi Cheng, Joseph A Moore, Michael Bowers, Ana P Kiess, Brandi R Page, Christine G Gourin TRM. An Oncospace Risk Prediction Model For Head And Neck Radiation Toxicities. AHNS 9th Int. Conf. Head Neck Cancer, Seattle, WA. 2016.http://ahns.jnabstracts.com/2016/Detail?ID=75998 (accessed 21 Dec 2017).

62 Lakshminarayanan P. *Radio-morphology: Parametric Shape-Based Features in Radiotherapy*. 2018.

63 Myronenko A, Xubo Song. Point Set Registration: Coherent Point Drift. *IEEE Trans Pattern Anal Mach Intell* 2010;**32**:2262–75. doi:10.1109/TPAMI.2010.46

64 Breiman L. Random forests. *Mach Learn* 2001;**45**:5–32. doi:10.1023/A:1010933404324

65 Lang JB. A Closer Look at Testing the 'No-Treatment-Effect' Hypothesis in a Comparative Experiment. *Stat Sci* 2015;**30**:352–71. doi:10.1214/15-STS513

66 Monti S, Palma G, D'Avino V, *et al.* Voxel-based analysis unveils regional dose differences

associated with radiation-induced morbidity in head and neck cancer patients. *Sci Rep* 2017;**7**:7220. doi:10.1038/s41598-017-07586-x

67 Nabi R, Shpitser I. Semi-Parametric Causal Sufficient Dimension Reduction Of High Dimensional Treatments. Published Online First: 2017.http://arxiv.org/abs/1710.06727

68 US. Environmental Protection Agency. Regulation of Fuels and Fuel Additives: Changes to Renewable Fuel Standard Program; Final Rule. *Fed Regist* 2010;**75**:14669–14904.https://www.gpo.gov/fdsys/pkg/FR-2010-03-26/pdf/2010-3851.pdf (accessed 27 Feb 2018).

69 US. Environmental Protection Agency. Final Renewable Fuel Standards for 2014, 2015 and 2016, and the Biomass-Based Diesel Volume for 2017. *Fed Regist* 2015;**80**:33100–52.https://www.epa.gov/renewable-fuel-standard-program/final-renewable-fuel-standards-2014-2015-and-2016-and-biomass-based (accessed 28 Feb 2018).

70 US. Energy Information Administration. November Monthly Energy Review, Table 7.3a Consumption of combustible fuels for electricity generation: Total (all sectors). https://www.eia.gov/totalenergy/data/monthly/index.php#electricity (accessed 28 Feb 2018).

71 The European Parliament and the Council of The European Union. Directive 2009/28/EC on the promotion of the use of energy from renewable sources and amending and subsequently repealing Directive 2001/77/EC and 2003/30/EC. *Off J Eur Union* 2009;:16–

62.http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32009L0028&from=EN (accessed 28 Feb 2018).

72 Food and Agriculture Organization of the United Nations. FAOSTAT. http://www.fao.org/faostat/en/#data/FT (accessed 28 Feb 2018).

73 Ehrig R, Behrendt F. Co-firing of imported wood pellets – An option to efficiently save CO2 emissions in Europe? *Energy Policy* 2013;**59**:283–300. doi:10.1016/J.ENPOL.2013.03.060

74 Xian H, Colson G, Mei B, *et al.* Co-firing coal with wood pellets for U.S. electricity generation: A real options analysis. *Energy Policy* 2015;**81**:106–16. doi:10.1016/J.ENPOL.2015.02.026

75 Lamers P, Schouwenberg PP, Marchal D, *et al.* Global wood chip trade for energy. 2012. http://task40.ieabioenergy.com/wp-content/uploads/2013/09/t40-global-wood-chips-study_final.pdf (accessed 28 Feb 2018).

76 Baral A, Malins C. COMPREHENSIVE CARBON ACCOUNTING FOR IDENTIFICATION OF SUSTAINABLE BIOMASS FEEDSTOCKS. 2014. https://www.theicct.org/sites/default/files/publications/ICCT_carbonaccounting-biomass_20140123.pdf (accessed 28 Feb 2018).

77 Nilsson LJ, Pisarek M, Buriak J, *et al.* Energy policy and the role of bioenergy in Poland. *Energy Policy* 2006;**34**:2263–78. doi:10.1016/J.ENPOL.2005.03.011

78 Paiano A, Lagioia G. Energy potential from residual biomass towards meeting the EU renewable energy and climate targets. The Italian case. *Energy Policy* 2016;**91**:161–73. doi:10.1016/J.ENPOL.2015.12.039

79 Trømborg E, Bolkesjø TF, Solberg B. Impacts of policy means for increased use of forest-based bioenergy in Norway—A spatial partial equilibrium analysis. *Energy Policy* 2007;**35**:5980–90. doi:10.1016/J.ENPOL.2007.08.004

80 Trømborg E, Havskjold M, Lislebø O, *et al.* Projecting demand and supply of forest biomass for heating in Norway. *Energy Policy* 2011;**39**:7049–58. doi:10.1016/J.ENPOL.2011.08.009

81 Lamers P, Junginger M, Hamelinck C, *et al.* Developments in international solid biofuel trade - An analysis of volumes, policies, and market factors. *Renew Sustain Energy Rev* 2012;**16**:3176–99. doi:10.1016/j.rser.2012.02.027

82 Goh CS, Junginger M, Cocchi M, *et al.* Wood pellet market and trade: a global perspective. *Biofuels, Bioprod Biorefining* 2013;**7**:24–42. doi:10.1002/bbb.1366

83 Junginger M, van Dam J, Zarrilli S, *et al.* Opportunities and barriers for international bioenergy trade. *Energy Policy* 2011;**39**:2028–42. doi:10.1016/J.ENPOL.2011.01.040

84 Ciner C, Dumas CF, Edward J, *et al. FINANCIAL ANALYSIS OF THE TRANSPORT OF WOOD CHIPS*. 2011.https://csb.uncw.edu/imba/annals/VonWuehlischC.pdf (accessed 28 Feb 2018).

85 Biotrade2020plus Approach to Sustainability, Report on the updated sustainability criteria to be considered for bioenergy for 2020 and 2030. 2016. http://www.biotrade2020plus.eu/images/BioTrade2020plus_D2.4_IINAS_Main_Report.pdf (accessed 28 Feb 2018).

86 Food and Agriculture Organization of the United Nations. FAOSTAT User'sGuide. http://faostat.fao.org/Portals/_Faostat/documents/pdf/FAOSTAT_User_Guide_v1_en.pdf (accessed 3 Sep 2016).

87 Hillring B. World trade in forest products and wood fuel. *Biomass and Bioenergy* 2006;**30**:815–25. doi:10.1016/j.biombioe.2006.04.002

88 Olsson O, Hillring B. The wood fuel market in Denmark – Price development, market efficiency and internationalization. *Energy* 2014;**78**:141–8. doi:10.1016/J.ENERGY.2014.09.065

89 Heinimö J. Methodological aspects on international biofuels trade: International streams and trade of solid and liquid biofuels in Finland. *Biomass and Bioenergy* 2008;**32**:702–16. doi:10.1016/J.BIOMBIOE.2008.01.003

90 Olsson O, Johnson FX. Bioenergy Trade in a Changing Climate -a review of mitigation-adaptation inter- relationships from a Nordic perspective. 2014. https://www.sei-international.org/mediamanager/documents/Publications/Climate/NORD-STAR-WP-2014-01-Olsson-Johnson.pdf (accessed 28 Feb 2018).

91 Samuelson PA. Spatial price equilibrium and linear programming. *Am Econ Rev* 1952;**42**:283–303. doi:10.2307/1810381

92 Takayama T, Judge GG. *Spatial and temporal price and allocation models.* North-Holland Pub. Co 1971. https://books.google.com/books/about/Spatial_and_temporal_price_and_allocatio.html?id=plNVAAAAMAAJ (accessed 28 Feb 2018).

93 Bawden DL. A Spatial Price Equilibrium Model of International Trade. *J Farm Econ* 1966;**48**:862. doi:10.2307/1236618

94 Lauri P, Havlík P, Kindermann G, *et al.* Woody biomass energy potential in 2050. *Energy Policy* 2014;**66**:19–31. doi:10.1016/J.ENPOL.2013.11.033

95 Kristöfel C, Strasser C, Morawetz UB, *et al.* Econometric analysis of the wood pellet market in Austria. 2015;:1–5.

96 Varian HR. *Intermediate Microeconomics: A Modern Approach*. W.W. Norton & Co 2010. doi:10.1017/CBO9781107415324.004

97 Golombek R, Gjelsvik E, Rosendahl KE. Effects of Liberalizing the Natural Gas Markets in Western Europe. *Energy J* 1995;**16**:85–111. doi:10.2307/41322587

98 Ahuja RK, Orlin JB. Inverse Optimization. *Oper Res* 2001;**49**:771–83. doi:10.1287/opre.49.5.771.10607

99 Chan TCY, Craig T, Lee T, *et al.* Generalized Inverse Multiobjective Optimization with

Application to Cancer Therapy. *Oper Res* 2014;**62**:680–95. doi:10.1287/opre.2014.1267

100 Jianzhong Zhang, Zhenhong Liu. Calculating some inverse linear programming problems. *J Comput Appl Math* 1996;**72**:261–73. doi:10.1016/0377-0427(95)00277-4

101 Howitt RE. Positive Mathematical Programming. *Am J Agric Econ* 1995;**77**:329. doi:10.2307/1243543

102 Heckelei T, Britz W. Models Based on Positive Mathematical Programming: State of the Art and Further Extensions. In: *The 89th EAAE Seminar*. Parma, Italy: 2005. 48–73.https://ageconsearch.umn.edu/bitstream/234607/2/Heckelei et al 2005 Models Based on Positive Mathematical Programming- State of the Art and Further Extensions.pdf (accessed 28 Feb 2018).

103 Bloomberg New Energy Finance. Next-generation ethanol and biochemicals: what's in it for Europe? 2010. https://www.dsm.com/content/dam/dsm/cworld/en_US/documents/bloomberg-nextgeneration-ethanol-and-biochemicals-whats-in-it-for-europe.pdf (accessed 28 Feb 2018).

104 Nahuelhual L, Carmona A, Lara A, *et al.* Land-cover change to forest plantations: Proximate causes and implications for the landscape in south-central Chile. *Landsc Urban Plan* 2012;**107**:12–20. doi:10.1016/J.LANDURBPLAN.2012.04.006

105 Vergara PM, Pérez-Hernández CG, Hahn IJ, *et al.* Deforestation in central Chile causes a

rapid decline in landscape connectivity for a forest specialist bird species. *Ecol Res* 2013;**28**:481–92. doi:10.1007/s11284-013-1037-x

106 Gaveau DLA, Linkie M, Suyadi, *et al.* Three decades of deforestation in southwest Sumatra: Effects of coffee prices, law enforcement and rural poverty. *Biol Conserv* 2009;**142**:597–605. doi:10.1016/J.BIOCON.2008.11.024

107 The Natural Resources Defense Council. Bioenergy Threatens the Heart of North American Wetland Forests. 2015. https://www.nrdc.org/sites/default/files/southeast-biomass-exports-FS.pdf (accessed 28 Feb 2018).

108 Biofuelwatch. Drax Plc lobbying of Government is misleading MPs and the public over biomass sustainability claims. 2013.http://www.biofuelwatch.org.uk/docs/Drax-PR.pdf (accessed 28 Feb 2018).

109 Correspondence regarding the financial loan provided to Drax Group plc to fund biomass and biofuels by the Green Investment Bank in 2012. 2013.https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/198127/eir-130334-correspondence-with-drax-group.pdf (accessed 28 Feb 2018).

110 Janssen R, Rutz DD. Sustainability of biofuels in Latin America: Risks and opportunities. *Energy Policy* 2011;**39**:5717–25. doi:10.1016/J.ENPOL.2011.01.047

111 Scarlat N, Dallemand J-F. Recent developments of biofuels/bioenergy sustainability certification: A global overview. *Energy Policy* 2011;**39**:1630–46.

doi:10.1016/J.ENPOL.2010.12.039

112 Lamers P, Hoefnagels R, Junginger M, *et al.* Global solid biomass trade for energy by 2020: an assessment of potential import streams and supply costs to North-West Europe under different sustainability constraints. *GCB Bioenergy* 2015;**7**:618–34. doi:10.1111/gcbb.12162

113 Siddiqui S, Christensen A. Determining energy and climate market policy using multiobjective programs with equilibrium constraints. *Energy* 2016;**94**:316–25. doi:10.1016/J.ENERGY.2015.11.002

114 Christensen A, Siddiqui S. Fuel price impacts and compliance costs associated with the Renewable Fuel Standard (RFS). *Energy Policy* 2015;**86**:614–24. doi:10.1016/J.ENPOL.2015.08.002

115 Christensen A, Siddiqui S. A mixed complementarity model for the US biofuel market with federal policy interventions. *Biofuels, Bioprod Biorefining* 2015;**9**:397–411. doi:10.1002/bbb.1545

116 Lutsey N, Sperling D. America's bottom-up climate change mitigation policy. *Energy Policy* 2008;**36**:673–85. doi:10.1016/J.ENPOL.2007.10.018

117 Sorda G, Banse M, Kemfert C. An overview of biofuel policies across the world. *Energy Policy* 2010;**38**:6977–88. doi:10.1016/J.ENPOL.2010.06.066

118 Berry T, Jaccard M. The renewable portfolio standard:: design considerations and an

implementation survey. *Energy Policy* 2001;**29**:263–77. doi:10.1016/S0301-4215(00)00126-9

119 Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2*. International Joint Conferences on Artificial Intelligence, Inc. 1995. 1137–43.https://dl.acm.org/citation.cfm?id=1643047 (accessed 5 Mar 2018).

120 Bengio Y. Gradient-Based Optimization of Hyper-Parameters. Published Online First: 1999.https://pdfs.semanticscholar.org/0b2b/806201fd1146f16b4c20fc3dd696e22c3c88.pdf (accessed 22 Feb 2018).

121 Do CB, Foo C-S, Ng AY. Efficient multiple hyperparameter learning for log-linear models. http://ai.stanford.edu/~chuongdo/papers/learn_reg.pdf (accessed 25 Jan 2018).

122 Pedregosa F. Hyperparameter optimization with approximate gradient. http://proceedings.mlr.press/v48/pedregosa16.pdf (accessed 29 Jan 2018).

123 Maclaurin D, Duvenaud D, Adams RP. Gradient-based Hyperparameter Optimization through Reversible Learning. https://arxiv.org/pdf/1502.03492.pdf (accessed 25 Jan 2018).

124 Fû J, Luô H, Fen J, *et al.* DrMAD: Distilling Reverse-Mode Automatic Differentiation for Optimizing Hyperparameters of Deep Neural Networks. https://arxiv.org/pdf/1601.00917.pdf (accessed 29 Jan 2018).

125 Jules MS. Experiments With Scalable Gradient-based Hyperparameter Optimization for

Deep Neural Networks. https://pdfs.semanticscholar.org/b694/d478bc9b7e4828c8fca4ff23433f9bf5e9d3.pdf (accessed 22 Feb 2018).

126 Franceschi L, Donini M, Frasconi P, *et al.* Forward and Reverse Gradient-Based Hyperparameter Optimization. https://arxiv.org/pdf/1703.01785.pdf (accessed 29 Jan 2018).

127 Hastie T, Edu H, Rosset S, *et al.* The Entire Regularization Path for the Support Vector Machine. *J Mach Learn Res* 2004;**5**:1391–415.http://delivery.acm.org/10.1145/1050000/1044706/hastie04a.pdf?ip=68.33.90.66&id=1044706&acc=OPEN&key=4D4702B0C3E38B35.4D4702B0C3E38B35.4D4702B0C3E38B35.6D218144511F3437&CFID=1016046698&CFTOKEN=65777655&__acm__=1514955666_48cca4fb11e80ca2ebe21df3ef88d3ad (accessed 2 Jan 2018).

128 Bennett KP, Kunapuli G, Hu J, *et al.* Bilevel optimization and machine learning. *Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics)* 2008;**5050 LNCS**:25–47. doi:10.1007/978-3-540-68860-0_2

129 Kunapuli G, Bennett KP, Hu J, *et al.* Bilevel Model Selection for Support Vector Machines. *Notes* 2008;**45**:129–58.

130 Kuhn HW, Tucker AW. NONLINEAR PROGRAMMING. In: *Proc. Second Berkeley Symp. on Math. Statist. and Prob.* 1951. 481–92. doi:10.1007/978-3-0348-0439-4

131 Couellan N, Wang W. Bi-level stochastic gradient for large scale support vector machine. *Neurocomputing* 2015;**153**:300–8. doi:10.1016/J.NEUCOM.2014.11.025

132 Shalev-Shwartz S, Singer Y, Srebro N, *et al.* Pegasos: primal estimated sub-gradient solver for SVM. *Math Program* 2011;**127**:3–30. doi:10.1007/s10107-010-0420-4

133 Zhang T, Tong. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In: *Twenty-first international conference on Machine learning - ICML '04*. New York, New York, USA: : ACM Press 2004. 116. doi:10.1145/1015330.1015332

134 Platt JC. Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines. *Adv kernel methods* 1998;:185–208. doi:10.1.1.43.4376

135 Boyd SP, Vandenberghe L. *Convex optimization*. Cambridge University Press 2004.

136 UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) Data Set. https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic) (accessed 7 Mar 2018).

137 UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/index.php (accessed 5 Mar 2018).

138 Chang C-C, Lin C-J. LIBSVM: A Library for Support Vector Machines. https://www.csie.ntu.edu.tw/~cjlin/papers/libsvm.pdf (accessed 15 Jan 2018).

139 UCI Machine Learning Repository. Connect-4 Data Set. http://archive.ics.uci.edu/ml/datasets/connect-4 (accessed 21 Mar 2018).

140 UCI Machine Learning Repository. MAGIC Gamma Telescope Data Set. 2007.https://archive.ics.uci.edu/ml/datasets/magic+gamma+telescope (accessed 21 Mar 2018).

141 McNutt, Wong J, Purdy J et al. OncoSpace: A new paradigm for clinical research and decision support in radiation oncology. In: *Proceedings of the XVIth International Conference on Computers in Radiation Therapy*. Amsterdam, Netherlands: 2010.

142 Goodfellow I, Bengio Y, Courville A. *Deep learning*. https://www.amazon.com/Deep-Learning-Adaptive-Computation-Machine/dp/0262035618/ref=sr_1_1?ie=UTF8&qid=1472485235&sr=8-1&keywords=deep+learning+book (accessed 8 Mar 2018).

143 Hindi H. A tutorial on optimization methods for cancer radiation treatment planning. *2013 Am Control Conf* 2013;:6804–16. doi:10.1109/ACC.2013.6580908

144 Buuren S van, Groothuis-Oudshoorn K. mice:Multivariate Imputation by Chained Equations in R. *J Stat Softw* 2011;**45**:1–67. doi:10.18637/jss.v045.i03

145 ROTNITZKY A, ROBINS JM. Semiparametric regression estimation in the presence of dependent censoring. *Biometrika* 1995;**82**:805–20. doi:10.1093/biomet/82.4.805

Curriculum Vitae

Wei Jiang was born in Wuhu, China on August 17, 1990. He received his Bachelor degree in Water Supply and Wastewater Engineering at Tongji University, Shanghai, China in 2012. After graduation from Tongji Univerisity, Wei enrolled in the Master of Science program in the Department of Geography and Environmental Engineering at Johns Hopkins University in Baltimore in fall 2012. Always interested in mathematics and physics, Wei chose the systems track during his master study. He has been taking courses in statistics, optimization, mathematical modeling, and machine learning since then. Wei researched water pipe network design using optimization methods under the supervision of Prof. Seth Guikema during his master study.

In June 2014, Wei started pursuing his Ph.D. degree in Civil Engineering at Johns Hopkins University. Through his Ph.D., he has focused on conducting researches in healthcare and energy systems applying machine learning and optimization methods. Wei has collaborated with multiple professors and researchers at Hopkins. Particularly, Prof. Scott Levin, Prof. Todd McNutt, and Dr. Harry Quon who are from the school of medicine at Johns Hopkins. He also worked closely with

Prof. Ilya Shpitser, Prof. Russ Taylor from the department of computer science at Johns Hopkins, and Prof. Sean Barnes from University of Maryland, College Park.

Wei was a teaching assistant for two classes at Johns Hopkins: Mathematical modeling, Probability and Statistics for Engineering. At the end of May, he will work as a data scientist at Staples, Inc. in Boston area.